MAXIMUM LIKELIHOOD ESTIMATION OF POISSON AND HAWKES PROCESSES AND EXTENSIONS TO HAWKES PROCESS ANALYSIS

A Dissertation
Presented to
The Academic Faculty

by

Michael George Moore

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Electrical and Computer Engineering

Georgia Institute of Technology

December 2018

MAXIMUM LIKELIHOOD ESTIMATION OF POISSON AND HAWKES PROCESSES AND EXTENSIONS TO HAWKES PROCESS ANALYSIS

Approved by:

Dr. Mark Davenport, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Eva Dyer
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Christopher Rozell
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Le Song
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Yao Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: August 14, 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF COMMONLY USED SYMBOLS AND OPERATORS

$\|\cdot\|_0$	number of nonzero entries in a vector or matrix
$\|\cdot\|_1$	ℓ_1 norm of vector or function or ℓ_1 induced norm of matrix
$\|\cdot\|_2$	ℓ_2 norm of vector or function or ℓ_2 induced norm (spectral norm) of matrix
$\|\cdot\|_\infty$	ℓ_∞ norm of vector or function or ℓ_∞ induced norm of matrix
$\|\cdot\|_{2,\infty}$	largest value of ℓ_2 norm of vector-valued function over its domain
$\|\cdot\|_{\max}$	largest-magnitude entry of matrix
$\|\cdot\|_F$	Frobenius norm of matrix
$\rho(\cdot)$	spectral radius of matrix
$\sigma(\cdot)$	N^{th}-largest singular value of $M \times N$ matrix
$\mathrm{Tr}(\cdot)$	trace of matrix
$\mathrm{diag}(\cdot)$	diagonal matrix with argument placed along diagonal
$[\cdot]_{rc}$	entry of argument at row r, column c
$\cdot^\top$	transpose of vector or matrix
$\preceq$	left argument minus right is negative semidefinite
$\succeq$	left argument minus right is positive semidefinite
$\mathbb{R}$, $\mathbb{R}_+$	set of real numbers, nonnegative real numbers
$\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{Z}_N$	set of integers, nonnegative integers, $\{1 \dots N\}$
$\|\cdot\|$	absolute value of scalar, measure of space, or cardinality of set
$\mathrm{supp}(\cdot)$	support of a vector (indices with nonzero entries)
$\mathcal{I}_{(\cdot)}$	indicator function that is 1 when the argument is true and 0 otherwise
$\mathbb{P}(\cdot)$	probability that argument is true
$\mathbb{E}(\cdot)$	expected value of random variable
Poisson$(\cdot)$	independent Poisson random variables with expectation of argument
Exponential$(\cdot)$	independent exponential random variables with expectation of argument
Bernoulli$(\cdot)$	independent Bernoulli random variables with expectation of argument

SUMMARY

The purpose of this work is to improve our ability to extract information from data generated by Poisson and Hawkes processes. Our principal focus is to provide improvements to the theory for parameter estimation under these models. To this end, we present novel bounds on the estimation of model parameters for linear Poisson processes. Earlier results applied to Poisson counting processes, but we improve upon these while addressing the broader class of Poisson arrival processes. For Hawkes processes, we present asymptotic parameter estimation bounds. These provide an enhanced understanding of how the structure of a Hawkes process affects our ability to estimate its parameters. We explore the capability of Hawkes processes to model telecommunication networks and to aid in the discovery of connections within such networks. We also discuss the considerations and extensions that should be employed when utilizing this model for this application. Finally, we discuss a way that the Hawkes process can be used to recognize event cascades within a network, rather than mere nodal connectivity. This is particularly relevant to telecommunication applications, as relay nodes often serve as intermediaries between others.

CHAPTER 1
INTRODUCTION

1.1 Motivation

The traditional sensing paradigm is based upon regularly or irregularly sampling some observable, such as voltage. Ideally, enough samples are taken such that no detail falls below the observation resolution. However, this method has its limits. Sometimes, as we seek finer and finer resolution, we run out of data to collect. For example, a functional MRI scan can be used to determine which parts of the brain are associated with different tasks. However, understanding the activity of the brain requires finer detail. Instead of bulk tissue with an average activity level, this finer detail gives us a collection of neurons with highly-stochastic activity and a limited number of spikes by-which to understand them.

This is related to the so-called "paradox of big data." The massive datasets we frequently encounter are often large collections of tiny data. As another example, a social media host can observe billions of interactions each day. From these, they can assemble detailed profiles of cities or countries. But the activity of an individual user may be limited to just a few interactions each day, and this restricts the precision of any individual assessment. Big data is capable of drawing broad inferences from vast collections of information, but small-scale assessments require a different set of tools.

The particular subset of small data with which we concern ourselves, here, is that of event-based data derived from point processes. These datasets arise when we consider information at the "atomic" scale. To recall our our earlier example, the bulk-averages of functional MRI become individual neuron firings at fine resolutions. Social media trends manifest as individual interactions at the personal level. At these scales, the things we observe do not follow Gaussian statistics because there are too-few observations for the

central limit theorem to be relevant. Whereas the dominant noise source in a radio receiver might be thermal noise in the amplifier or external interference, the dominant noise source in our examples is the variability inherent to individual observable events.

There are many different point processes used to model many different kinds of events, but we will limit our discussion to just two.

The first point process we will consider is the ubiquitous Poisson process. Poisson processes work on the principle of independence between events and form a useful model for a large range of event-based applications. Among such applications are nuclear imaging [1], fluoroscopy [2], mass spectrometry [3], low-light imaging [4–6], and medical image processing [7]. The simplicity of Poisson models makes them the preferred choice over other point processes when their independence assumption can be (adequately) met.

The second process we will consider is the Hawkes process. Unlike Poisson processes, Hawkes processes are autoregressive in that actions can create reactions. This allows for feedback within the system. Hawkes processes can, for example, be used to model the relationship between earthquakes and aftershocks [8]. Other applications of Hawkes processes include modeling social networks [9, 10], communication networks [11], neural networks [12], financial transactions [13], ecology [14], and sociological patterns [15–17].

In order to make the most of our limited samples, it is useful to know just how representative they are. In other words, we would like to understand how precise a model we can create by fitting parameters to our available data. Improvements in this understanding would aid in the responsible use of the inferred models. While this question possesses limited answers for Poisson processes, there are almost no such results for Hawkes processes. The core motivation of this work is to expand upon both of these bodies of literature. The secondary aim of this work is to expand upon the use of Hawkes models in a few practical respects, with special interest given to their use in modeling telecommunication networks.

1.2 Contributions

The first contribution of this work is to extend parameter recovery guarantees for Poisson counting processes to the more-general class of Poisson arrival processes. We present this work in Chapter 2 and discuss the improvements that this allows. We then provide asymptotic recovery bounds for Hawkes processes in Chapter 3. Such results represent an important step in understanding the behavior of Hawkes processes and, in particular, placing the appropriate degree of confidence in Hawkes models learned from data. Motivated by our particular interest in wireless radio networks, Chapter 4 offers a number of adaptions and experiments that consider the use of Hawkes processes in this application. Lastly, in Chapter 5 we explore the use of Hawkes processes to attribute the anonymous target of a broadcast interaction. This allows us to ascribe event-level detail and interpretations to observations. It is particularly useful in the radio network application, where complex relaying schemes can otherwise limit the relevance of Hawkes processes.

1.3 Background

1.3.1 Poisson processes

We will begin our discussion with Poisson processes. Poisson processes are the simplest example of a point process and form the base case for more-sophisticated models.

First we must define a Poisson random variable. A Poisson random variable, which we denote $Y \sim \text{Poisson}(\lambda)$ for parameter y, takes values over the nonnegative integers $\mathbb{Z}_+$ and has probability mass function

$$\mathbb{P}(Y = y) = \frac{\lambda^y}{y!} \exp(-\lambda) \quad y \in \mathbb{Z}_+. \tag{1.1}$$

Note that the expectation and variance $\mathbb{E}(Y) = \mathbb{E}((Y - \mathbb{E}(Y))^2) = \lambda$. This stands in contrast to a Gaussian random variable, for example, where the signal (expectation) and noise

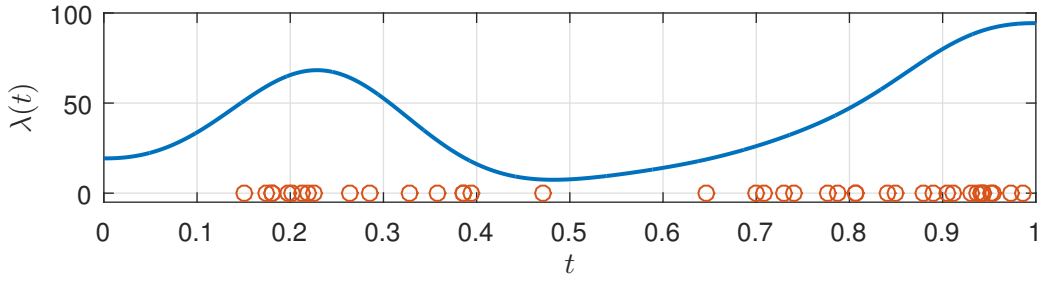

Figure 1.1: Example realization of a point process. We observe event locations (circles) but their empirical density provides only a sketch of the underlying intensity (line).

(variance) are individually specified. For Poisson observations, this inability to decouple signal from noise renders many powerful and well-studied estimation techniques inapplicable, demanding an entirely separate set of tools. For this reason, much of the analysis we present will depart from familiar techniques used to treat Gaussian noise.

Poisson random variables form the underpinnings of point processes. We will consider point process over a domain $\mathbb{T}$ (e.g., time, space, etc.) that we sample according to an intensity (also called a density or rate) function $\lambda : \mathbb{T} \to \mathbb{R}_+$. A point process manifests as a countable set of *events*, indexed by S (e.g., a set of integers), with coordinates $\tau_k \in \mathbb{T}$ for each $k \in S$. We define our set of all coordinates $\tau_S = \{\tau_m\}_{m \in S}$. The events satisfy the relationship[1]

$$|\tau_S \cap T| \sim \text{Poisson}\left(\int_T \lambda(t)dt\right) \tag{1.2}$$

for any $T \subseteq \mathbb{T}$. In words, the expected number of events over any interval is proportional to the integrated intensity over that interval and the probability of an event occurring at any particular instant is proportional to the instantaneous intensity.

An example realization of a Poisson process is presented in Figure 1.1. As with any small collection of samples, the empirical density gives only a coarse approximation of the true intensity. Unfortunately, this is the regime in which we typically operate when handling point processes. We cannot rely on the samples, alone, to provide a high-quality

[1]Depending on context, we will use the $|\cdot|$ operator to denote the cardinality of a set, the measure of a space, or the absolute value of a scalar.

estimate of an unknown intensity function. Instead, we will find it necessary to compensate for the limited availability of data with model assumptions.

The negative log-likelihood of a set of event coordinates τ_S under intensity $\lambda(t)$ is

$$\mathcal{L}(\tau_S|\lambda) = \int_{\mathbb{T}} \lambda(t)dt - \sum_{m\in S} \log \lambda(\tau_m). \tag{1.3}$$

Henceforth, we will refer to the number of events observed as the quantity $M = |S|$. Typically, we will define the indices $S = \{1 \dots M\}$. For a Poisson process, $M \sim \text{Poisson}(\int_{\mathbb{T}} \lambda(t)dt)$. We will also find it convenient to define the expected value $\overline{M} = \mathbb{E}(M) = \int_{\mathbb{T}} \lambda(t)dt$.

The existence of a negative log-likelihood leads us to a natural way to estimate λ. If we assume that the intensity function lies in some set $\lambda \in \mathcal{R}$ then the maximum likelihood estimator is

$$\widehat{\lambda} = \arg\min_{\lambda\in\mathcal{R}} \mathcal{L}(\tau_S|\lambda). \tag{1.4}$$

If $\lambda(t)$ is parameterized by some set of variables x, as a notational shorthand we will often replace the above objective with $\mathcal{L}(\tau_S|x)$ and the likelihood-maximizing parameters with $\widehat{x}$.

If $\mathcal{R} = \{\lambda : \lambda(t) \geq 0 \ \forall t \in \mathbb{T}\}$ then the maximum likelihood estimate is trivially given by $\widehat{\lambda}(t) = \sum_{m=1}^{M} \delta(t - \tau_m)$, where $\delta(t)$ is the Dirac delta function. However, this is a degenerate explanation that results in a gross over-fitting of the model. Because we usually expect the rate to be supported more broadly, it is necessary to impose some degree of regularity on $\mathcal{R}$. This is commonly done via a generalized linear model (GLM).

The simplest GLM for point process data, and the one that will be our focus, is the *linear model*

$$\mathcal{R} = \{\lambda \in \text{span}(\{\gamma_1 \dots \gamma_N\}) : \lambda(t) \geq 0 \ \forall t \in \mathbb{T}\} \tag{1.5}$$

where each $\gamma_n : \mathbb{T} \to \mathbb{R}$. This admits a representation parameterized by $x \in \mathbb{R}^N$,

$$\lambda_x(t) = \langle x, \gamma \rangle = \sum_{n=1}^{N} x_n \gamma_n(t). \tag{1.6}$$

If we define $b \in \mathbb{R}^N$ and $A \in \mathbb{R}^{M \times N}$ as $b_n = \int_{\mathbb{T}} \gamma_n(t) dt$ and $A_{mn} = \gamma_n(\tau_m)$ then we can rewrite the negative log-likelihood as

$$\mathcal{L}(\tau_S | x) = b^\top x - 1^\top \log(Ax) \tag{1.7}$$

where 1 is the vector with each entry equal to one and the logarithm is taken element-wise over the vector result of Ax. Substituting this likelihood into (1.4) yields a convex program that is solvable via a variety of possible techniques. Some specially-adapted solvers for this particular program are described in [18, 19] and many more address extensions of this problem, usually by including some form of additional penalization on certain solutions such as in [9].

An extremely useful property of Poisson random variables is that they are superimposable [20]. By this, we mean that if $X \sim \text{Poisson}(x)$ and $Y \sim \text{Poisson}(y)$ are independent Poisson random variables and we define another random variable $Z = X + Y$, then it is also true that $Z \sim \text{Poisson}(x + y)$. A consequence of this result is that if a set of events $\{\theta\}$ are drawn from a point process with intensity $\lambda_\theta(t)$ and events $\{\psi\}$ are drawn from an independent point process with intensity $\lambda_\psi(t)$ then the union of these events $\{\{\theta\} \cup \{\psi\}\}$ could equivalently have been drawn from a point process with intensity $\lambda_\theta(t) + \lambda_\psi(t)$ [21].

A Poisson process is the simplest variety of point process. Its defining property is that the intensity function fixed. In other varieties, the intensity function can instead develop and evolve, often via a historical dependence on itself.

1.3.2 Counting processes

So far, we have described an *arrival process*. An arrival process measures the exact coordinate of each individual event. An alternative form of point process is a *counting process*. A counting process is realized by histogramming an arrival process into nonoverlapping bins, corresponding to a $\mathbb{T}$ that represents some countable set. This model is preferred for sensors with limited resolution, such as particle or photon detector arrays (e.g., [1]). Instead of observing arrival coordinates, we observe the number of arrivals in each predetermined bin. For bin m with support T_m, the bin count y_m has distribution

$$y_m \sim \text{Poisson}\left(\int_{T_m} \lambda(t)dt\right). \tag{1.8}$$

Bins are assumed to be nonoverlapping so that all observations are independent.

We will refer to the number of bins as $M_0 = \dim(y)$. If we assume a linear representation, as in (1.6), we can write

$$\int_{T_m} \lambda_x(t)dt = \sum_{n=1}^{N} x_n \int_{T_m} \gamma_n(t)dt = \sum_{n=1}^{N} x_n A_{mn} \tag{1.9}$$

by defining matrix $A \in \mathbb{R}^{M_0 \times N}$ with $A_{mn} = \int_{T_m} \gamma_n(t)dt$. The negative log-likelihood is

$$\mathcal{L}(y|x) = 1^\top Ax - y^\top \log(Ax). \tag{1.10}$$

If $\gamma_n(t)$ is constant over each bin, the counting model and arrival model are functionally identical. However, results for the two are not completely interchangeable. While one might hope to adapt counting process results into arrival process applications, this is rarely possible. Driving the bin sizes to zero (to approximate the nondiscrete arrival model) results in a vanishing number of events within each interval. The vanishing counts typically produce vacuous results when used with existing theory, as discussed in [22]. Because arrival processes generalize counting processes, we will frame our discussion in terms of

arrival processes when possible.

1.3.3 Hawkes processes

A (univariate) Hawkes process is a point process that exhibits a specific autoregressive behavior. In a Hawkes process, the intensity function is a function of the preceding events and is given by

$$\lambda(t|\tau_S) = \mu(t) + \sum_{k \in S} \gamma(t - \tau_k). \tag{1.11}$$

This is often referred to as the *conditional intensity function* to emphasize its dependence on the events τ_S. The *base intensity function* $\mu(t)$ is nonnegative. The *excitation kernel* $\gamma(t)$ is integrable, nonnegative, and strictly causal (i.e., $\gamma(t) = 0$ for all $t \leq 0$) and dictates the autoregressive behavior of the process. The fact that the excitation kernel is strictly causal means that we only need to consider the preceding events (with $\tau_k < t$) to determine the present intensity; present or future events have no effect on the present or past intensity.

A multivariate Hawkes process generalizes a Hawkes process to a collection of N point processes. The intensity of each constituent *subprocess* is a function of previous events on all subprocesses. We continue to use S to index the set of all events, but partition S into N sets S_i for $i \in \mathbb{Z}_N$ (define $\mathbb{Z}_N = \{1, 2, \ldots N\}$). The subset S_i represents the indices of all events arising on subprocess i, so that $k \in S_i$ indexes an event on subprocess i. The intensity of subprocess i, indexed by $i \in \mathbb{Z}_N$, has the form

$$\lambda_i(t|\tau_S) = \mu_i(t) + \sum_{j \in \mathbb{Z}_N} \sum_{k \in S_j} \gamma_{ij}(t - \tau_k). \tag{1.12}$$

We consider all subprocesses taken together to form the multivariate Hawkes process. We say that an event k is *type-j* if $k \in S_j$. Again, the excitation kernels $\gamma_{ij}(t)$ are integrable, nonnegative, and strictly causal. We will define the *excitation matrix*[2] to have entries $A_{ij} =$

[2]Which is not to be confused with the sampled basis matrix for the Poisson process, discussed in Section 1.3.1, which shares the same symbol.

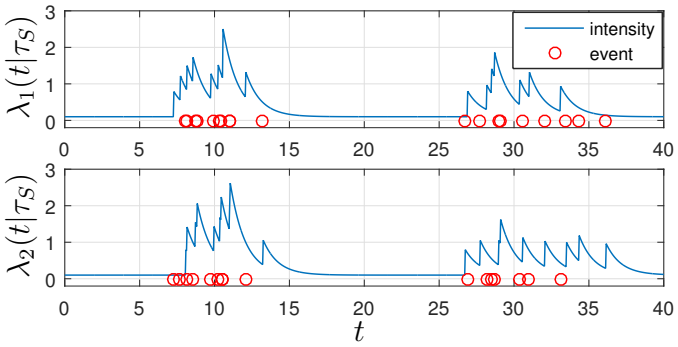

Figure 1.2: Intensity functions and events for a bivariate Hawkes process.

$\int \gamma_{ij}(t)dt$ and the *total base intensity vector* to have entries $u_i = \int \mu_i(t)dt$. We assume both quantities to be finite.

In Figure 1.2, we show an example of a multivariate Hawkes process with two subprocesses. In this example, the base intensity vectors are $\mu_1(t) = \mu_2(t) = 0.1$ and the kernels are decaying exponentials. We use $A_{12} = A_{21} = 0.7$ and $A_{11} = A_{22} = 0$. Given this excitation matrix, each subprocess excites the other (but not itself), causing the two to form clusters of events in the same regions.

We will assume that the spectral radius of the excitation matrix, denoted by $\rho(A)$, satisfies $\rho(A) < 1$. As we will discuss in Chapter 3, violation of this condition allows for the possibility that $\lim_{t\to\infty} \lambda_i(t) = \infty$, which makes the process both difficult to observe (in that we cannot record an infinite number of events) and breaks most correspondence with reality. We will use the term *Hawkes process* to refer to both univariate and multivariate Hawkes processes.

Like the Poisson process, nonlinear variants of Hawkes-like processes exist. While some simply rectify the intensity function [23], others apply logarithmic nonlinearities [24, 25]. Others seek to learn the nonlinear link function from the data itself [26].

1.3.4 Linear autoregressive point processes

The discretized version of a Hawkes process is sometimes referred to as a linear autoregressive point process (LARPP). In an LARPP, we can write our observations as a sequence in

$t \in \mathbb{Z}_+$ of N-vectors such that the event count

$$y_t \sim \text{Poisson}\left(\mu_t + \sum_{k=1}^{\infty} G_k y_{t-k}\right) \tag{1.13}$$

where $G_k \in \mathbb{R}_+^{N \times N}$ (for $k \in \{1, 2, \ldots\}$) and Poisson(v) represents a vector of independent Poisson random variables with expectation v.

To see the connection to multivariate Hawkes processes, suppose that the functions $\mu_i(t)$ and $\gamma_{ij}(t)$ are only nonzero at regularly-spaced (without loss of generality, integer) values of t. To be specific, the LARPP corresponds to a Hawkes process with $\mu(t) = \sum_k \mu_k \delta(t-k)$ and $\gamma_{ij}(t) = \sum_k [G_k]_{ij} \delta(t-k)$. The observations y_t are mapped $[y_t]_i = \sum_{k \in S_i} \mathcal{I}_{(\tau_k = t)}$, where $\mathcal{I}_{(\cdot)}$ is 1 if the argument is true and zero otherwise. As with other Hawkes processes, we will assume that the excitation matrix $A = \sum_{k=1}^{\infty} G_k$ satisfies $\rho(A) < 1$.

Because the LARPP is a special case, we will direct the majority of our analysis at the more-general Hawkes process.

1.3.5 Marked Hawkes processes

Adding *marks* to a Hawkes process provides a method to tag events with additional information that may impact their contribution to the conditional intensity function. We will discuss this in more detail in Chapter 4, but for now we will only introduce a simple case.

It is possible to consider treat a multivariate Hawkes process as a marked univariate Hawkes process. To do this, we add marks $\theta_k \in \mathbb{Z}_N$ for $k \in S$. These marks replace the subsets S_j that we discussed earlier. With these new mark variables, we can write the intensity (1.12) as

$$\lambda(t|\tau_S, \theta_S) = \sum_{i \in \mathbb{Z}_N} \mu_i(t) + \sum_{j \in \mathbb{Z}_N} \sum_{k \in S} \mathcal{I}_{(\theta_k = j)} \gamma_{ij}(t - \tau_k). \tag{1.14}$$

For a marked Hawkes process, it is also necessary to define a *mark generating func-*

tion that describes the distribution of marks at a given event coordinate. For our multivariate Hawkes process, we define the subintensity functions $\breve{\lambda}_i(t|\tau_S, \theta_S) = \mu_i(t) + \sum_{j\in\mathbb{Z}_N}\sum_{k\in S}\mathcal{I}_{(\theta_k=j)}\gamma_{ij}(t-\tau_k)$ and the mark distribution function is

$$\mathbb{P}(\theta_k = i|\tau_k) = \frac{\breve{\lambda}_i(\tau_k|\tau_S, \theta_S)}{\sum_{j\in\mathbb{Z}_N}\breve{\lambda}_j(\tau_k|\tau_S, \theta_S)} \tag{1.15}$$

While the dependence on the set θ_S (which would include θ_k) appears problematic, remember that $\gamma_{ij}(t)$ is zero for nonpositive arguments. As such, it has no impact on itself (or future events) and can be generated independently. To convert the realization (τ_S, θ_S) to a multivariate Hawkes process, set $S_j = \{k \in S : \theta_k = j\}$.

1.3.6 Simulating Poisson processes

When simulating a counting model, one does not need to sample events. One can simply generate the number of points in each bin as a Poisson random variable with the specified intensity. Simulating an arrival process is slightly more complicated, but is easily accomplished using rejection sampling, sometimes called *thinning*.

Rejection sampling works by taking a point process with a larger rate and then discarding events to achieve the lower rate. Suppose that $\lambda'(t)$ is an intensity function satisfying $\lambda'(t) \geq \lambda(t)$ and τ' are events drawn independently from a Poisson process with intensity $\lambda'(t)$. If we keep event τ'_i with probability $\lambda(\tau'_i)/\lambda'(\tau'_i)$ and discard it otherwise, then the nondiscarded events are equivalently drawn from a Poisson process with intensity $\lambda(t)$.

In practice, we typically choose $\lambda'(t) = \lambda_{\text{max}}$ where $\lambda_{\text{max}} = \sup_{t\in\mathbb{T}}\lambda(t)$. This homogeneous Poisson process (having a constant intensity) can be generated by choosing $M' = \text{Poisson}(|\mathbb{T}|\lambda_{\text{max}})$ and then choosing τ' to be M' coordinates sampled uniformly from $\mathbb{T}$. We then perform the rejection step, as described, to be left with our simulated points τ corresponding to intensity $\lambda(t)$.

If the intensity function has a large dynamic range (is very small over some regions and

very large over others), it is possible to subdivide $\mathbb{T}$ and use a λ_{max} over each region that more-precisely bounds the intensity. This can accelerate simulation by reducing the number of rejected samples. It is also possible to generate a Poisson process (in a one-dimensional space) using the method described in Section 1.3.7.

1.3.7 Simulating Hawkes processes

To simulate Hawkes processes, we will employ a more elaborate version of rejection sampling. The method we will describe is referred to as Ogata's thinning algorithm [27]. It is a form of rejection sampling where we generate observations sequentially instead of as a batch.

Suppose, without loss of generality, that our Hawkes process uses $\mathbb{T} = [0, T]$. We begin by considering the earliest coordinate, $q = 0$. We then generate the next arrival of a Poisson process with a too-high intensity. For the too-high intensity, use $r = \sup_{t \in (q,T]} \lambda(t|\tau)$. In a Poisson process, the interval between subsequent events is exponentially distributed. Accordingly, we generate an exponential random variable $s \sim \text{Exponential}(r)$ and update q to be $q + s$, the new candidate coordinate. If the new $q > T$, we are finished. If we are not finished, we add the coordinate to τ with probability $\lambda(q|\tau)/r$ and reject it otherwise. If we keep the sample and the process is marked, we also generate its accompanying mark according to the mark distribution function at its location. We then repeat the process by computing a new intensity supremum and considering another candidate.

By only generating the single next event, we can continuously update the intensity function to account for its historical dependence. A slight modification of this algorithm considers only a finite future window at a time, which can be helpful if the future intensity function grows substantially in the not-immediate future (which would otherwise lead to many rejections before reaching the higher intensity).

Alternatively, one can generate Hawkes processes by using their status as a branching process. For a further discussion of Hawkes processes as branching processes, see Chap-

ter 3.

Simulating a LARPP is similar to the simulation of Poisson counting processes. One can simply generate the number of events in the next bin as a Poisson random variable with the appropriate expectation, without the need for considering the intricacies of arrival processes.

CHAPTER 2

LINEAR POISSON PROCESS ESTIMATION

We consider the estimation of linear parameterizations of Poisson processes with intensity functions of the form (1.6). We will particularly focus on arrival processes, which, despite their status as a generalization of counting processes, receive comparatively little attention. In doing so, we provide novel arrival process parameter estimation guarantees.

Despite the expanded generality, compared to results addressing only counting processes, our contribution reflects an improvement over existing results in some regimes. We also present a regularization scheme that can overcome the theoretical difficulties of low-intensity processes before providing evidence that these apparent difficulties may be an illusory artifact of the analysis.

2.1 Related work

The majority of research involving parametric Poisson estimation is based on counting processes, as in, e.g., [22, 28–32]. Recall that these observations take the form $y \sim \text{Poisson}(Ax)$ for a known A. The expectation $\mathbb{E}(y) = Ax$ suggests a naive estimation program based on solving the least-squares program $\widehat{x} = \arg\min_x \|y - Ax\|_2$. However, this estimator is poorly suited to Poisson recovery because of the non-Gaussian nature of y, its signal-dependent variance in particular. A popular alternative (see, e.g., [1]) suggests instead using a weighted least-squares program that solves $\min_x \|W(y - Ax)\|_2$ subject to some basic constraints and application-specific penalties. In this case, W is used to normalize the data-dependent variance across different terms in y.

Recent interest in sparse signal processing has lead researchers to suggest a weighted LASSO [29], which augments the least squares objective with a sparsity-inducing ℓ_1-

regularization term. Their objective function takes the form

$$\widehat{x} = \arg\min_{x} \|y' - A'x\|_2 + \beta\|Dx\|_1, \tag{2.1}$$

where y' and A' are modified versions of their namesakes and D is a set of carefully-chosen weights. This merges weighted least-squares methods with a sparsity assumption on x, where β is used to modulate the strength of this prior. Recovery guarantees, bounding the difference between the estimated and true x, exist for this program under certain assumptions. [30] instead uses an unweighted LASSO for this task. Another LASSO-based estimator is considered in [33], but this result is rare in that it is also applicable to arrival processes. However, again it provides bounds only on the accuracy with which the intensity is estimated and does not yield parameter estimation guarantees.

A major shortcoming of the weighted least-squares approaches to Poisson estimation is that they require priors or hyperparameters to help determine, oftentimes only approximately, the optimal weights. Maximum likelihood estimators operate directly on the likelihood function, removing the need for hyperparameters. Recall that a Poisson process likelihood takes the form of (1.7) or, for a counting process, (1.10). The corresponding MLE is a convex program and [18, 19] provide effective programs for performing the optimization (also allowing for regularization of the parameters, similarly to in LASSO).

Despite its practical advantages, the Poisson MLE is challenging to analyze. It does not benefit from the extensive efforts that have gone toward analyzing least-squares estimators. Only recently have explicit guarantees been made for the Poisson MLE. In particular, [32] provides a recovery guarantee bounding $\|\widehat{x} - \bar{x}\|_2$ under the Poisson counting model when A and x are nonnegative and shows that its bound is order-optimal in several model parameters. It also handles the case when x is sparse.

The work of [22, 34] provide upper and lower bounds on the best possible estimators under the Poisson linear model with some additional constraints (sparsity, nonnegativity,

flux preservation) using carefully-constructed measurement matrices. While valuable for comparison, these bounds do not provide actionable guarantees for specific recovery programs.

Performance bounds for constrained or penalized maximum-likelihood estimators are provided in [22, 31, 32]. Since our results also imply bounds for the counting model, it is instructive to compare our bounds with the existing work in this domain. Among the prior work which is most relevant is that of [32], from which we specifically draw inspiration in portions of our analysis. A detailed comparison of our results with those in [32] is provided in Section 2.8.

2.2 Maximum likelihood parameter estimation

We will introduce a slight extension of the model introduced in (1.6). We will introduce a baseline intensity function $g : \mathbb{T} \to \mathbb{R}$ and define

$$\lambda_x(t) = g(t) + \sum_{n=1}^{N} x_n \gamma_n(t). \tag{2.2}$$

This addition allows us to consider a known component of the intensity function. Similarly, we update the negative log-likelihood (1.7) to

$$\mathcal{L}(\tau_S|x) = b^\top x - 1^\top \log(g + Ax) \tag{2.3}$$

with $g_m = g(\tau_m)$. Note that A and g are both random, as they depend on the random observations τ_S.

Suppose that the true parameter values of a process are $\bar{x}$. Further, suppose that we know that $\bar{x} \in \mathcal{X}$ for some set $\mathcal{X}$. At the very least, we know that $\bar{x} \in \{x : \lambda_x \in \mathcal{R}\}$. The

maximum likelihood estimate of the parameters is

$$\widehat{x} = \arg\min_{x \in \mathcal{X}} \mathcal{L}(\tau_S | x). \tag{2.4}$$

For convex $\mathcal{X}$, the estimation program (2.4) is convex. This optimization problem can be efficiently solved by algorithms such as those presented in [18, 19].

2.3 Parameter estimation error bound

In order to state our parameter estimation result, we will make use of the definition $\Gamma = \text{Gram}(\gamma(t))$, i.e., the $N \times N$ Gram matrix with entries $\Gamma_{ij} = \int_{\mathbb{T}} \gamma_i(t)\gamma_j(t)dt$. Let $\text{Tr}(\Gamma)$ and $\sigma(\Gamma)$ represent the trace and minimum eigenvalue of Γ, respectively, and let $\|\gamma\|_{2,\infty} = \sup_{t \in \mathbb{T}} \|\gamma(t)\|_2$.

With these definitions, we can state our result as follows:

Theorem 1 *If events τ_S are produced by a Poisson arrival process with intensity $\lambda_{\bar{x}}(t) \in \{\{0\} \cup [\lambda_{\text{min}}, \lambda_{\text{max}}]\}$ and $\lambda_{\text{min}} > 2\zeta \frac{\|\gamma\|_{2,\infty}^2}{\sigma(\Gamma)}$, any vector $\widehat{x}$ satisfying $\lambda_{\widehat{x}}(t) \in [0, \lambda_{\text{max}}]$ and $\mathcal{L}(\tau_S|\widehat{x}) \leq \mathcal{L}(\tau_S|\bar{x})$ will also satisfy*

$$\|\widehat{x} - \bar{x}\|_2 < c \frac{\sqrt{\zeta \,\text{Tr}(\Gamma)}}{\sigma(\Gamma)} \frac{\lambda_{\text{max}}}{\sqrt{\lambda_{\text{min}}}} \tag{2.5}$$

for a specific constant c with probability at least $1 - (2N+1)\exp(-\zeta)$.

The proof is detailed in Section 2.12.

We note that $g(t)$ has no impact on the result of Theorem 1. This opens the door to what seems like a rather unorthodox strategy for improving this bound. Specifically, the dependence on the dynamic range $\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$ can be removed by simply augmenting our observations with additional events ρ_V generated by a homogeneous Poisson process with intensity λ_{max}. These additional events will increase the intensity of our process, but in a

manner completely known to us. Specifically, in this case the intensity of $\tau_S \cup \rho_V$ is

$$\widetilde{\lambda}_{\bar{x}}(t) = \lambda_{\bar{x}}(t) + \lambda_{\text{max}}. \tag{2.6}$$

Note that if $\lambda_{\bar{x}}(t) \in [0, \lambda_{\text{max}}]$ then $\widetilde{\lambda}_{\bar{x}}(t) \in [\lambda_{\text{max}}, 2\lambda_{\text{max}}]$ and the corresponding $\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \leq 2$. Thus, by simply replacing $g(t)$ with $g(t) + \lambda_{\text{max}}$ in (2.3) and applying Theorem 1, we obtain the following corollary:

Corollary 1 *If events τ_S are produced by a Poisson arrival process with intensity $\lambda_{\bar{x}}(t) \in [0, \lambda_{\text{max}}]$, events ρ_V are produced by a homogeneous Poisson process with intensity λ_{max}, and $\lambda_{\text{max}} > 2\zeta \frac{\|\gamma\|_{2,\infty}^2}{\sigma(\Gamma)}$, then any vector $\widehat{x}$ satisfying $\lambda_{\widehat{x}}(t) \in [0, \lambda_{\text{max}}]$ and $\mathcal{L}(\tau_S \cup \rho_V | \widehat{x}) \leq \mathcal{L}(\tau_S \cup \rho_V | \bar{x})$ will also satisfy*

$$\|\widehat{x} - \bar{x}\|_2 < 2c \frac{\sqrt{\zeta \operatorname{Tr}(\Gamma)}}{\sigma(\Gamma)} \sqrt{\lambda_{\text{max}}} \tag{2.7}$$

for a specific constant c with probability at least $1 - (2N+1)\exp(-\zeta)$.

To clarify, when we augment our observations τ_S with the set of events ρ_V (with known intensity β), the negative log-likelihood function that results is

$$\mathcal{L}(\tau_S \cup \rho_V | x) = \int_{\mathbb{T}} \lambda_x(t) dt - \sum_{m \in S} \log(\beta + \lambda_x(\tau_m)) - \sum_{\ell \in V} \log(\beta + \lambda_x(\rho_\ell)). \tag{2.8}$$

Comparing the bound in (2.7) with that in (2.5), we see that in exchange for an additional factor of 2, we have completely removed the dependence on the dynamic range. Moreover, we have replaced the constraint on λ_{min} in Theorem 1 with a more relaxed constraint on λ_{max}. Nevertheless, one might be justifiably skeptical that this will lead to improved performance in practice – we are adding a *large* amount of noise to the signal we wish to estimate. We explore this empirically in Section 2.10. Here, we note that a potentially more palatable strategy might be to perform a deterministic form of regularization

that is similar in spirit. Specifically, define the regularized negative log-likelihood

$$\mathcal{L}'_\beta(\tau_S|x) = \int_{\mathbb{T}} \lambda_x(t)dt - \sum_{m\in S} \log\left(\beta + \lambda_x(\tau_m)\right) - \beta \int_{\mathbb{T}} \log\left(\beta + \lambda_x(t)\right) dt, \tag{2.9}$$

noting that this is equivalent to (2.3) when $\beta = 0$. $\mathcal{L}'_\beta(\tau_S|x)$ is the result of replacing $\sum_\ell \log(\beta + \lambda_x(\rho_\ell))$ from $\mathcal{L}(\tau_S \cup \rho_V|x)$ with its expectation.[1]

Unfortunately, there does not appear to be an easy extension of our current analysis techniques that would allow us to provide an error bound for the estimate obtained by minimizing (2.9). We conjecture that a result similar to Corollary 1 should also hold when optimizing the regularized negative log-likelihood in (2.9) with $\beta = \lambda_{\text{max}}$.

2.4 Estimation under sparsity assumptions

In many practical settings, especially when the dimension N of the basis $\gamma(t)$ is large, one might expect $\bar{x}$ to be sparse (i.e., to have relatively few nonzeros). There is a natural extension of Theorem 1 that can result in an improved bound when this occurs. To state this result, we will define $\gamma_U(t)$ to be the $|U|$-dimensional basis obtained by taking the basis elements $\gamma_n(t)$ indexed by U and Γ_U to be the $|U| \times |U|$ submatrix composed of the rows and columns of Γ indexed by U (i.e., the Gram matrix of $\gamma_U(t)$). Let $\text{Tr}(\Gamma_U)$ and $\sigma(\Gamma_U)$ represent the trace and minimum eigenvalue of Γ_U, respectively, and let $\|\gamma_U\|_{2,\infty} = \sup_{t\in\mathbb{T}} \|\gamma_U(t)\|_2$.

Theorem 2 *If events τ_S are produced by a Poisson arrival process with intensity $\lambda_{\bar{x}}(t) \in \{\{0\} \cup [\lambda_{\text{min}}, \lambda_{\text{max}}]\}$ and $\lambda_{\text{min}} > 2\zeta \frac{\|\gamma_U\|^2_{2,\infty}}{\sigma(\Gamma_U)}$, any vector $\widehat{x}$ satisfying $\text{supp}(\widehat{x} - \bar{x}) \subseteq U$,*

[1]The difficulty of computing the integral in (2.9) in closed-form means that, typically, numerical integration would be necessary when performing estimation using the regularized likelihood in (2.9). The use of Riemann integration reduces this result to a sampled version of an arrival model, i.e., a counting model. Recovery as in Corollary 1 does not suffer this drawback, remaining comparable to standard maximum likelihood estimation as featured in Theorem 1.

$\lambda_{\widehat{x}}(t) \in [0, \lambda_{\text{max}}]$, *and* $\mathcal{L}(\tau_S|\widehat{x}) \leq \mathcal{L}(\tau_S|\bar{x})$ *will also satisfy*

$$\|\widehat{x} - \bar{x}\|_2 < c\frac{\sqrt{\zeta \operatorname{Tr}(\Gamma_U)}}{\sigma(\Gamma_U)}\frac{\lambda_{\text{max}}}{\sqrt{\lambda_{\text{min}}}} \tag{2.10}$$

for a specific constant c *with probability at least* $1 - (2|U| + 1)\exp(-\zeta)$.

The proof is detailed in Section 2.12. Because it follows as a trivial analogue of the non-sparse version, we omit the formal statement of the sparse version of Corollary 1, but naturally a similar improvement (removing dependence on the dynamic range) is again possible here. We (again) conjecture that similar improvements should be possible using a deterministic regularizer.

Note that the bound in Theorem 2 depends on the specific choice of index set U. A reader familiar with the sparse approximation literature may wonder how this compares with more traditional results, which hold for arbitrary sparsity patterns, and how the quantities $\operatorname{Tr}(\Gamma_U)$ and $\sigma(\Gamma_U)$ compare to more familiar quantities. In fact, it is straightforward to restate this result in terms of an appropriate analogue of the *restricted isometry property* (RIP) [35]. Specifically, let $\mathcal{W} : \mathbb{R}^N \to L_2(\mathbb{T})$ denote the operator defined by $\mathcal{W}(x) = \langle x, \gamma \rangle$. Observe that we can write $\|\mathcal{W}(x)\|^2_{L_2(\mathbb{T})} = x^\top \Gamma x$. We say that the operator $\mathcal{W}$ (or equivalently, the basis $\gamma(t)$) satisfies the RIP if there exists a constant $\delta_s \in [0, 1)$ such that

$$(1 - \delta_s)\|x\|_2^2 \leq x^\top \Gamma x \leq (1 + \delta_s)\|x\|_2^2 \tag{2.11}$$

for all x with at most s nonzeros. Using this definition, if we have $\operatorname{supp}(\widehat{x} - \bar{x}) \subseteq U$ as in Theorem 2, where $|U| \leq s$, then from the RIP we have the bounds $\operatorname{Tr}(\Gamma_U) \leq s(1 + \delta_s)$ and $\sigma(\Gamma_U) \geq 1 - \delta_s$. Accordingly, the bound (2.10) from Theorem 2 can be replaced with

$$\|\widehat{x} - \bar{x}\|_2 < c\frac{\sqrt{\zeta s(1 + \delta_s)}}{1 - \delta_s}\frac{\lambda_{\text{max}}}{\sqrt{\lambda_{\text{min}}}}. \tag{2.12}$$

We note that this substitution is compatible with the regularization scheme of Corollary 1,

which would remove the dynamic range term from (2.12).

Observe that the results of Theorem 2 (as stated above or combined with bounds based on the RIP) are really only of interest when we can ensure that the set $\mathrm{supp}(\widehat{x} - \bar{x})$ is sufficiently small. If we assume that $\bar{x}$ has only s nonzeros, then we can ensure this provided that we can guarantee that $\widehat{x}$ is also sparse. This is trivially true if we constrain $\widehat{x}$ to have at most s nonzeros by encoding this in the constraint set $\mathcal{X}$ in the optimization problem of (2.4), in which case $|\mathrm{supp}(\widehat{x} - \bar{x})| \leq 2s$. Unfortunately, this leads to a more challenging nonconvex optimization problem. An open question concerning (2.4) is whether we can guarantee that $\widehat{x}$ is sparse when using a convex sparsity-inducing constraint such as $\mathcal{X} = \{x \in \mathbb{R}^N : \|x\|_1 \leq \eta, \lambda_x \in \mathcal{R}\}$.

2.5 Guarantees for counting processes

When using the Poisson counting model (1.8), the special cases of Theorem 1 and Corollary 1 follow immediately from the discussion in Section 1.3.2. Specifically, if we define $\|A\|_{2,\infty} = \max_m \sqrt{\sum_n A_{mn}^2}$, $\|A\|_F$ to be the Frobenius norm of A, and $\sigma(A)$ to be the smallest singular value of A (the minimum eigenvalue of $\sqrt{A^\top A}$), then we have the following result:

Corollary 2 *Let $y \sim$ Poisson$(g + A\bar{x})$ with $g + A\bar{x} \in \{\{0\} \cup [\lambda_{\min}, \lambda_{\max}]\}^{M_0}$. If $\lambda_{\min} > 2\zeta \frac{\|A\|_{2,\infty}^2}{\sigma^2(A)}$, any vector $\widehat{x}$ satisfying $g + A\widehat{x} \in [0, \lambda_{\max}]^{M_0}$ and $\mathcal{L}(y|\widehat{x}) \leq \mathcal{L}(y|\bar{x})$ will also satisfy*

$$\|\widehat{x} - \bar{x}\|_2 < c \frac{\sqrt{\zeta}\|A\|_F}{\sigma^2(A)} \frac{\lambda_{\max}}{\sqrt{\lambda_{\min}}} \tag{2.13}$$

for a specific constant c with probability at least $1 - (2N + 1)\exp(-\zeta)$.

For the counting process version of Corollary 1, we again augment our observations with uniform Poisson data with intensity $\lambda_{\max}$ and replace g with $g + \lambda_{\max}$. This substitution yields the following improved bound:

Corollary 3 *Let $y \sim$ Poisson$(g + A\bar{x})$ with $g + A\bar{x} \in [0, \lambda_{\max}]^{M_0}$ and $\rho \in \mathbb{R}^{M_0}$ be a vector of independent random variables with $\rho_m \sim$ Poisson$(\lambda_{\max})$. If $\lambda_{\max} > 2\zeta \frac{\|A\|_{2,\infty}^2}{\sigma^2(A)}$, any vector $\widehat{x}$ satisfying $g + A\widehat{x} \in [0, \lambda_{\max}]^{M_0}$ and $\mathcal{L}(y + \rho|\widehat{x}) \leq \mathcal{L}(y + \rho|\bar{x})$ will also satisfy*

$$\|\widehat{x} - \bar{x}\|_2 < 2c \frac{\sqrt{\zeta}\|A\|_F}{\sigma^2(A)} \sqrt{\lambda_{\max}} \tag{2.14}$$

for a specific constant c with probability at least $1 - (2N + 1)\exp(-\zeta)$.

We again conjecture that a similar result to Corollary 3 should also hold when using the regularized negative log-likelihood, which in this case would take the form of $\mathcal{L}'_\beta(y|x) = 1^\top Ax - (\beta + y)^\top \log(\beta + g + Ax)$.

Finally, we note that these results can easily be extended to the sparse setting in a manner analogous to Theorem 2 by replacing Γ with Γ_U, $\gamma(t)$ with $\gamma_U(t)$, A with A_U, N with $|U|$, and adding the constraint $\operatorname{supp}(\widehat{x} - \bar{x}) \subseteq U$.

2.6 Sample complexity

A common lens through which we can evaluate a statistical estimation problem concerns the number of observations required to obtain an accurate estimate of the quantities of interest. Here our situation is somewhat outside the standard framework as the parameters $\bar{x}$ which we wish to estimate actually determine the number of observations (events) we will obtain. Nevertheless, we can gain some insight by approaching the problem from this perspective.

For the purpose of generality, we will analyze the number of observations with respect to Theorem 2. In the dense case of Theorem 1, simply set $U = \{1, \ldots, N\}$. It is necessary that $\zeta > \log(2|U|+1)$ for the bound to hold with nonzero probability. Because $|U|\sigma(\Gamma_U) \leq \operatorname{Tr}(\Gamma_U) \leq |\mathbb{T}|\|\gamma_U\|_{2,\infty}^2$, the requirement that $\lambda_{\min} > 2\zeta \frac{\|\gamma_U\|_{2,\infty}^2}{\sigma(\Gamma_U)}$ implies that it is necessary that $|\mathbb{T}|\lambda_{\min} \gtrsim |U| \log |U|$ for the bound to hold with nonzero probability. Alternatively, in the regularized case of Corollary 1 we replace the requirement on $\lambda_{\min}$ with a requirement

on λ_{max}, resulting in the condition that $|\mathbb{T}|\lambda_{\text{max}} \gtrsim |U| \log |U|$. On the surface, neither of these conditions directly addresses the issue of how many observations are required, but we note that $|\mathbb{T}|\lambda_{\text{min}} \leq \overline{M} \leq |\mathbb{T}|\lambda_{\text{max}}$. Thus when the dynamic range is modest, these conditions essentially reduce to $\overline{M} \gtrsim |U| \log |U|$. This mirrors typical results in the standard setting of sparse recovery with Gaussian statistics [35].

2.7 Choosing a constraint set

If $\bar{x} \in \mathcal{X}$ then the program in (2.4) will result in an estimate $\widehat{x}$ that, by construction, satisfies the $\mathcal{L}(\tau_S|\widehat{x}) \leq \mathcal{L}(\tau_S|\bar{x})$ condition of our theorems. Possible sets $\mathcal{X}$ can include unconstrained vectors, sparse vectors, vectors with limited ℓ_p norms, or nonnegative vectors. Any convex $\mathcal{X}$ will maintain the convexity of (2.4) – although convexity is not necessary if we are not concerned with computational efficiency or can still ensure that $\mathcal{L}(\tau_S|\widehat{x}) \leq \mathcal{L}(\tau_S|\bar{x})$ through some alternative argument that does not rely on solving (2.4) to global optimality. This results in a tradeoff between constraint sets $\mathcal{X}$ that enable simple efficient algorithms and $\mathcal{X}$ that more tightly enforce desired properties in our recovery $\widehat{x}$. For example, when sparsity is exploited (as in Theorem 2) the set $\mathcal{X} = \{x : \|x\|_0 \leq \|\bar{x}\|_0\}$ (where $\|x\|_0$ counts the number of nonzeros in x) is a natural choice to ensure a fixed sparsity level in $\widehat{x}$, but results in a nonconvex (and thus challenging) optimization problem. In contrast, $\mathcal{X} = \{x : \|x\|_1 \leq \eta\}$ is convex and encourages sparse solutions, but does not necessarily guarantee a specific sparsity level.

We also note that our Theorems depend on the condition that $\lambda_{\widehat{x}}(t) \in [0, \lambda_{\text{max}}]$. This can be satisfied by choosing $\mathcal{X} \subseteq \{x : 0 \leq \lambda_x(t) \leq \lambda_{\text{max}}\}$. While the exact value of λ_{max} will likely be unknown, it can be estimated or bounded to sufficient accuracy from τ_S. Alternatively, it can be bounded as $\lambda_{\text{max}} \leq \|\gamma\|_{2,\infty}\|\bar{x}\|_2$ when we have knowledge of the norm of $\bar{x}$, although this bound is typically poor. As a matter of practice, we believe that there is often minimal risk in ignoring this constraint completely. Its role is solely to limit the pointwise difference $\sup_{t\in\mathbb{T}} |\lambda_{\bar{x}}(t) - \lambda_{\widehat{x}}(t)| \leq \lambda_{\text{max}}$, but we expect that a suitable

$\widehat{x}$ will accomplish this (at least up to a constant factor) even in the absence of an explicit constraint.

2.8 Comparison to existing results

Here we summarize the result of [32] in terms of the quantities introduced here. The result holds when A, $\widehat{x}$, and $\bar{x}$ are constrained to be nonnegative and $\|\bar{x}\|_0 \leq s$ holds for some choice s. To make the comparison, we will need to provide several additional definitions from [32] (relabeled to resemble our notation):

$$\sigma_*^2(A) = \min_{\substack{U:|U|\leq s,\\ x:\|x_U\|_1 \geq \|x - x_U\|_1}} \frac{\|Ax\|_2^2}{\|x\|_2^2}$$

$$\lambda_{\text{min}}^* = \min_{\substack{x:|x\|_0 = s,\\ \|x\|_1 = \|\bar{x}\|_1}} H(g + Ax)$$

$$\lambda_{\text{max}}^* = \|g\|_\infty + \|A\|_{\text{max}} \|\bar{x}\|_1$$

where $H(\cdot)$ denotes the harmonic mean of the input vector and $\|\cdot\|_{\text{max}}$ denotes the maximum-magnitude entry of a matrix. The quantity $\sigma_*^2(A)$ denotes the restricted eigenvalue of A of order s and is roughly comparable to $\min_{|U|\leq s} \sigma^2(A_U)$. Additionally, λ_{min}^* is roughly on the order of λ_{min}. Finally, note that $\lambda_{\text{max}}^* \geq \lambda_{\text{max}}$ and $\sqrt{sM_0}\|A\|_{\text{max}} \geq \|A_U\|_F$ (for any $|U| \leq s$) and that these inequalities are often quite loose.

With these definitions, the main result of [32] can be written as

$$\|\widehat{x} - \bar{x}\|_2 \leq 54 \left(3 + \log \tfrac{\lambda_{\text{max}}^*}{\lambda_{\text{min}}}\right) \frac{\sqrt{\zeta s M_0}\|A\|_{\text{max}}}{\sigma_*^2(A)} \frac{\lambda_{\text{max}}^*}{\sqrt{\lambda_{\text{min}}^*}} \tag{2.15}$$

with probability at least $1 - 2\exp(-\zeta)$ when $\zeta \leq \frac{M_0 \lambda_{\text{min}}}{H(g + A\bar{x})} \min\{1, \lambda_{\text{min}}\}$. We remark that this is extremely similar to the findings of the sparse variant of Corollary 2. However, Corollary 2 removes the restriction that A, $\widehat{x}$, and $\bar{x}$ be nonnegative. Further, Theorem 2 generalizes this result to the case of Poisson arrival processes. We have also presented

Corollary 1, which removes the dependence on $\lambda_{\min}$ or $\lambda^*_{\min}$ entirely. Although it was not the primary aim of this work, we have also considerably improved the dependence on the terms $\lambda^*_{\max}$ and $\|A\|_{\max}$. However, we also stress that the results in [32] do have an advantage when dealing with sparse $\bar{x}$ in that their guarantees hold for the (convex) ℓ_1 constrained estimate (as opposed to requiring the nonconvex sparsity constraint required to obtain a similar guarantee via Theorem 2).

2.9 Optimality

Here we briefly discuss the optimality of our results. In particular, we will first consider a concrete choice of $\gamma(t)$ for which the expected performance is easy to directly calculate and compare this performance to the bounds established above. Next, we will characterize the Cramér-Rao lower bound for this problem and compare our results with this bound.

2.9.1 Example: A simple orthobasis

Consider the simple orthobasis of

$$\gamma_n(t) = \begin{cases} 1 & n-1 \leq t < n \\ 0 & \text{otherwise.} \end{cases}$$

In this case we have $\Gamma = I$ and $\|\gamma(t)\|_{2,\infty} = 1$. Since this basis is disjoint (no two elements are supported on intersecting intervals), this is really just a collection of independent Poisson estimation problems with $y_m \sim \text{Poisson}(\bar{x}_m)$. The maximum likelihood estimate is trivially given by

$$\widehat{x}_n = \sum_m \gamma_n(\tau_m),$$

i.e., the number of events falling in that interval. Hence, $\widehat{x}_n \sim \text{Poisson}(\bar{x}_n)$ and the expected squared error is $\mathbb{E}(\widehat{x}_n - \bar{x}_n)^2 = \bar{x}_n$, from which it follows that

$$\mathbb{E}\|\widehat{x} - \bar{x}\|_2^2 = \sum_{n=1}^{N} x_n.$$

Applying Theorem 1 to this problem yields the bound

$$\|\widehat{x} - \bar{x}\|_2^2 \leq c^2 \zeta N \lambda_{\text{max}} \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$$

with probability at least $1-(2N+1)\exp(-\zeta)$ when $\lambda_{\text{min}} > 2\zeta$. Our high-probability bound differs from the average-case error of this system (up to a constant) only by the dynamic range. Alternatively, in Corollary 1, we showed how the dynamic range could be removed from our bound by using a modified recovery program. Specifically, we obtain a guarantee of the form

$$\|\widehat{x} - \bar{x}\|_2^2 \leq 4c^2 \zeta N \lambda_{\text{max}}$$

with probability at least $1 - (2N+1)\exp(-\zeta)$ when $\lambda_{\text{max}} > 2\zeta$. In this case, the bound differs from the expectation by only constants and the ratio between the average and maximum intensity $\bar{x}_n$.

2.9.2 Cramér-Rao lower bound

The example described above suggests that for a well-conditioned basis $\gamma(t)$, our analysis appears to be relatively tight. Indeed, by comparing our bound to the Cramér-Rao lower bound for this problem, we will obtain additional evidence that this is, in fact, the case.

Towards this end, we note that the Fisher information matrix for (1.7) is

$$\mathfrak{I}_x = \mathbb{E}\left(\nabla_x^2 \mathcal{L}(\tau_S|x)\right) = \mathbb{E}\left(A^T \operatorname{diag}(g + Ax)^{-2} A\right).$$

The Cramér-Rao lower bound states that any unbiased estimate $\widehat{x}$ of $\bar{x}$ must satisfy

$$\mathbb{E}\|\widehat{x} - \bar{x}\|_2^2 \geq \mathrm{Tr}(\mathfrak{I}_{\bar{x}}^{-1}).$$

The precise value of $\mathrm{Tr}(\mathfrak{I}_{\bar{x}}^{-1})$ will depend on the problem (on both $\gamma(t)$ and on $\bar{x}$). However, one can show that $\mathbb{E}\left(A^T \operatorname{diag}(g + A\bar{x})^{-1} A\right) = \Gamma$ (see Section 2.12.2), which provides the semidefinite ordering

$$\Gamma^{-1}\lambda_{\min} \preceq \mathfrak{I}_{\bar{x}}^{-1} \preceq \Gamma^{-1}\lambda_{\max}.$$

From this we obtain

$$\mathrm{Tr}(\Gamma^{-1})\lambda_{\min} \leq \mathrm{Tr}(\mathfrak{I}_{\bar{x}}^{-1}) \leq \mathrm{Tr}(\Gamma^{-1})\lambda_{\max}. \tag{2.16}$$

Because the maximum-likelihood estimator is asymptotically efficient [36] (achieves the Cramér-Rao bound for asymptotically large sets of observations), (2.16) suggests that (asymptotically) our MLE will satisfy a bound of the form

$$\mathbb{E}\|\widehat{x} - \bar{x}\|_2^2 \leq \mathrm{Tr}(\Gamma^{-1})\lambda_{\max} \tag{2.17}$$

Note that some improvements may be possible in special cases (e.g., in the example of Section 2.9.1 we can replace $\lambda_{\max}$ with the *mean* intensity). However, there is limited room for improvement here since (2.16) also implies that a bound of the form

$$\mathbb{E}\|\widehat{x} - \bar{x}\|_2^2 \leq \mathrm{Tr}(\Gamma^{-1})\lambda_{\min}$$

is impossible, as it would violate the Cramér-Rao bound.

Compare these bounds to the result of Corollary 1, which states that (with high probability) we can achieve

$$\|\widehat{x} - \bar{x}\|_2^2 \lesssim \frac{\mathrm{Tr}(\Gamma)}{\sigma(\Gamma)^2}\lambda_{\max}. \tag{2.18}$$

Contrasting (2.17) with (2.18), we observe that (2.17) is a slightly stronger guarantee (ignoring the fact that (2.18) holds with high probability while (2.17) holds only in expectation). In particular, it is a straightforward consequence of the eigendecomposition of Γ that

$$\mathrm{Tr}(\Gamma^{-1}) \leq \frac{\mathrm{Tr}(\Gamma)}{\sigma(\Gamma)^2},$$

with equality if and only if the spectrum of Γ is flat (i.e., all eigenvalues are equal to each other). In the case where the basis is ill-conditioned and Γ has one or more eigenvalues which are very small compared to the remainder, the bound in (2.17) can be somewhat tighter than that in (2.18). However, we note that unless N is quite large, both $\mathrm{Tr}(\Gamma^{-1})$ and $\frac{\mathrm{Tr}(\Gamma)}{\sigma(\Gamma)^2}$ can both be dominated by $\frac{1}{\sigma(\Gamma)}$, and so when the basis is very poorly conditioned both bounds are rather poor. In total, these results suggest that (up to a constant), there is relatively little room for improvement over the bound in Corollary 1.

2.10 Regularization in practice

Here we take a practical look into the claims of Corollary 1 and our conjecture concerning our deterministically regularized alternative. Specifically, Corollary 1 provides improved theory (at least when λ_{min} is small) when we augment our observations with additional random events (i.e., noise). We would like to understand if such gains are actually to be expected in practice.

Towards this end, we choose a basis of $N = 50$ shifted-Gaussian basis functions, sampled as to produce a counting process with $M_0 = 500$ observations ($A \in \mathbb{R}^{M_0 \times N}$). We choose coefficient vectors $\bar{x}$ with entries that are independent and identically exponentially distributed such that the expected number of events is $\overline{M} = 100$. We then simulate the Poisson process with the intensity given by this basis and these parameters and compute the ℓ_2-norm error between the true $\bar{x}$ and maximum likelihood estimate $\widehat{x}$. We compare this error to the error of computing the regularized estimator using random (Corollary 1) or

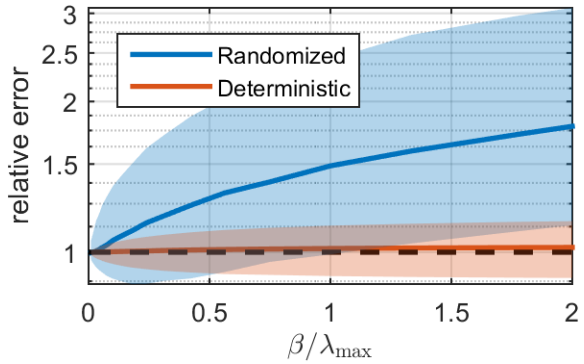

Figure 2.1: Comparison of estimation error $\|\widehat{x} - \bar{x}\|_2$ with regularization to error without regularization. Lines represent the median relative error and the shaded region denotes the space between the 10% and 90% quantiles. Values less than one denote improvement over unregularized recovery while values above one denote degraded performance.

deterministic (using (2.9)) regularization. Figure 2.1 plots this result for different choices of regularization parameter β relative to $\lambda_{\max}$ over many trials. Note that Corollary 1 uses $\beta = \lambda_{\max}$, and larger β denotes stronger regularization (and $\beta = 0$ denotes no regularization). Relative error larger than 1 denotes degraded performance with regularization and less than 1 denotes improved performance.

While we only include simulations for a single choice of basis and parameters, these results are representative of the behavior we observed for every choice of basis/parameters we have attempted. Examining Figure 2.1, we note that randomized regularization (Corollary 1) frequently results in significantly-degraded performance. Larger regularization results in larger error as the noise overwhelms the signal. Deterministic regularization does not suffer the same increasing error with increasing regularization, but instead saturates at error that is, in the median, slightly higher than the unregularized error. We infer that the effect of regularization is simply to perturb the solution in a way that will sometimes improve the error and will sometimes degrade it, without biasing our solution in some way that results in consistent improvement.

In summary, despite our best efforts, we have not uncovered any regime in which the regularization suggested by Corollary 1 (or our conjecture) systematically improved recovery over the unregularized case. This leaves open the possibility that the bound of

Corollary 1 may be more representative of actual performance than that of Theorem 1 (differing by a dependence on the dynamic range of the underlying intensity) even when the prescribed regularization is completely omitted. In other words, it is possible that the results of Corollary 1 actually hold without any regularization at all. However, our analysis does not provide an obvious means to establish this result.

2.11 Discussion

In this chapter we have provided novel recovery guarantees for parameter estimation in the context of Poisson arrival processes. These guarantees are near-optimal. However, we note that this optimal theoretical performance is established for a version of the MLE which has been augmented with noise – a procedure that does not appear to actually result in improved performance in practice.

There are several remaining improvements that may yet be possible. First, the results presented here require that (the unknown part of) the Poisson intensity function lie within the span of the basis. It may also be possible to address recovery when the true intensity is only well-approximated by this basis. Second, the existence of Corollary 1, which adds noise to the observations to improve theoretical performance, combined with the findings of Sections 2.9 and 2.10 suggests that the dynamic range dependence in Theorem 1 may be an artifact of the proof and ultimately unnecessary. We continue to conjecture that an improved bound should be possible for our deterministically-regularized MLE (or a similar alternative), but leave further exploration of this issue for future work. Finally, many important open questions remain concerning the use of more sophisticated convex regularizers (e.g., an ℓ_1-norm constraint when $\bar{x}$ is sparse).

2.12 Proofs

The overall outline of our proof is inspired by the proof of the main result of [32]. We will provide a proof of Theorem 2. Theorem 1 follows as a special case of the same argument.

Define $\Delta' = \widehat{x} - \bar{x}$. Our goal is to show that

$$\|\Delta'\|_2 < \epsilon \qquad \forall \Delta' \text{ s.t. } \mathcal{L}(\tau_S|\bar{x} + \Delta') < \mathcal{L}(\tau_S|\bar{x}). \tag{2.19}$$

A sufficient condition for (2.19) is that

$$\mathcal{L}(\tau_S|\bar{x} + \Delta') - \mathcal{L}(\tau_S|\bar{x}) \geq 0 \qquad \forall \Delta' \text{ s.t. } \|\Delta'\|_2 \geq \epsilon. \tag{2.20}$$

In other words, any large deviation from $\bar{x}$ implies an inferior likelihood to that of $\bar{x}$, and hence any $\widehat{x}$ with a smaller negative log-likelihood than $\bar{x}$ must lie close to $\bar{x}$.

Suppose that Δ' has s nonzero elements and $\mathrm{supp}(\Delta') = U$. We can define a $\Delta \in \mathbb{R}^s$ and a selector matrix $J \in \{0,1\}^{N \times s}$ (a submatrix of the $N \times N$ identity matrix) such that $\Delta' = J\Delta$. We make the definitions $b_U = J^T b$, $A_U = AJ$, $\Gamma_U = J^T \Gamma J$, and $\gamma_U(t) = J\gamma(t)$. As a shorthand, we will define $D = \mathrm{diag}(g + A\bar{x})^{-1}$. The negative log-likelihood gap between the perturbed and true solutions reduces to

$$\mathcal{L}(\tau_S|\bar{x} + J\Delta) - \mathcal{L}(\tau_S|\bar{x}) = b_U^T \Delta - 1^T \log\left(1 + DA_U\Delta\right).$$

We use the bound $\log(1+x) \leq x - \frac{3x^2}{6+4x}$, along with some simple algebraic manipulation, to obtain

$$\mathcal{L}(\tau_S|\bar{x}+J\Delta) - \mathcal{L}(\tau_S|\bar{x}) \geq b_U^T\Delta - 1^T DA_U\Delta + 3\Delta^T A_U^T D\,\mathrm{diag}(6(g+A\bar{x}) + 4A_U\Delta)^{-1} A_U\Delta. \tag{2.21}$$

Constraining $\lambda_{\bar{x}}(t)$ and $\lambda_{\widehat{x}}(t)$ to the interval $[0, \lambda_{\max}]$ ensures that $\|g + A\bar{x}\|_\infty \leq \lambda_{\max}$ and $\|A_U\Delta\|_\infty \leq \lambda_{\max}$. Employing these bounds allows us state that

$$\mathrm{diag}(6(g + A\bar{x}) + 4A_U\Delta)^{-1} \succeq \frac{1}{10\lambda_{\max}} I$$

and thus we can lower-bound the right-hand side of (2.21) by

$$(b_U - A_U^T D1)^T \Delta + \frac{3}{10\lambda_{\max}} \Delta^T A_U^T D A_U \Delta. \tag{2.22}$$

To satisfy (2.20), it is sufficient to show that (2.22) is nonnegative for all $\|\Delta\|_2 \leq \epsilon$. After applying the Cauchy-Schwarz inequality and basic properties of singular values, we obtain

$$(2.22) \geq -\left\|b_U - A_U^T D1\right\|_2 \|\Delta\|_2 + \frac{3\sigma(A_U^T D A_U)}{10\lambda_{\max}} \|\Delta\|_2^2,$$

where $\sigma(\cdot)$ denotes the minimum singular value of matrix. Accordingly,

$$\|\Delta\|_2 \geq \frac{10\lambda_{\max} \left\|b_U - A_U^T D1\right\|_2}{3\sigma(A_U^T D A_U)} \tag{2.23}$$

implies that $\mathcal{L}(\tau_S|\bar{x} + J\Delta) - \mathcal{L}(\tau_S|\bar{x}) \geq 0$. This means that the choice

$$\epsilon = 10\lambda_{\max} \left\|b_U - A_U^T D1\right\|_2 / 3\sigma(A_U^T D A_U)$$

is sufficient to satisfy (2.20). This expression involves two random quantities, $\|b_U - A_U^T D1\|_2$ and $\sigma(A_U^T D A_U)$. These are random because A and D depend on the (random) observations τ_S of our process. We address these quantities using the following lemmas:

Lemma 1 *Using the preceding definitions, the inequality*

$$\left\|b_U - A_U^T D1\right\|_2 \leq \frac{2}{3} \frac{\zeta \|\gamma_U\|_{2,\infty}}{\lambda_{\min}} + \sqrt{2\zeta \frac{M}{\overline{M}} \frac{\mathrm{Tr}(\Gamma_U)}{\lambda_{\min}}}$$

holds with probability at least $1 - (s+1)\exp(-\zeta)$. ***Proof: See Section 2.12.1.***

Lemma 2 *Using the preceding definitions, the inequality*

$$\sigma(A_U^T D A_U) \geq \sigma(\Gamma_U) \left(1 - \sqrt{\frac{2\zeta \|\gamma_U\|_{2,\infty}^2}{\lambda_{\min} \sigma(\Gamma_U)}}\right)$$

holds with probability at least $1 - s\exp(-\zeta)$. *Proof: See Section 2.12.2.*

By applying Lemmas 1 and 2 to our expression for ϵ and applying a union bound, with probability at least $1 - (2s+1)\exp(-\zeta)$ the error is bounded

$$\epsilon \leq \frac{10\lambda_{\max}}{3} \frac{\frac{2}{3}\frac{\zeta\|\gamma_U\|_{2,\infty}}{\lambda_{\min}} + \sqrt{2\zeta\frac{\mathrm{Tr}(\Gamma_U)}{\lambda_{\min}}}}{\sigma(\Gamma_U)\left(1 - \sqrt{\frac{2\zeta\|\gamma_U\|_{2,\infty}^2}{\lambda_{\min}\sigma(\Gamma_U)}}\right)}.$$

If $\lambda_{\min} \geq \alpha\zeta\frac{\|\gamma_U\|_{2,\infty}^2}{\sigma(\Gamma_U)}$ for some α then, using the inequalities $s\sigma(\Gamma_U) \leq \mathrm{Tr}(\Gamma_U) \leq |\mathbb{T}|\|\gamma_U\|_{2,\infty}^2$, we bound

$$\zeta \leq \frac{\lambda_{\min}\sigma(\Gamma_U)}{\alpha\|\gamma_U\|_{2,\infty}^2} \leq \frac{\lambda_{\min}\,\mathrm{Tr}(\Gamma_U)}{\alpha s\|\gamma_U\|_{2,\infty}^2}. \tag{2.24}$$

By defining $c_{\alpha,s} = \frac{10}{3}\frac{\frac{2}{3\sqrt{\alpha s}}+\sqrt{2}}{1-\sqrt{\frac{2}{\alpha}}}$, it follows that

$$\epsilon \leq c_{\alpha,s}\frac{\sqrt{\zeta\,\mathrm{Tr}(\Gamma_U)}}{\sigma(\Gamma_U)}\frac{\lambda_{\max}}{\sqrt{\lambda_{\min}}} \tag{2.25}$$

with probability at least $1 - (2s+1)\exp(-\zeta)$.

The constant $c_{\alpha,s}$ shrinks as α or s grow (though other terms in the bound grow with s). For any $\alpha > 2$, $c_{\alpha,s}$ is well-defined and bounded by an absolute constant for any s. For example, with the trivial choice $s \geq 1$ we can bound $c_{3,s} < 33$, $c_{5,s} < 16$, and $c_{\infty,s} \approx 4.71$. This completes the proof of Theorem 2. Theorem 1 is a result of the choice $s = N$ (i.e., $J = I$).

2.12.1 Proof of Lemma 1

The elements of the vector $b - A^T D1$ are given by the expression

$$\left[b - A^T D1\right]_n = b_n - \sum_{m=1}^{M}\frac{\gamma_n(\tau_m)}{\lambda_{\bar{x}}(\tau_m)}.$$

Recall that the event coordinates τ_m are independent and identically distributed with probability density function $f_{\tau_m}(t) = \frac{\lambda_{\bar{x}}(t)}{\overline{M}}$ and that $M \sim \text{Poisson}(\overline{M})$. Define a positive integer $Q \geq \overline{M}$, and $w_q^{(Q)} \sim \text{Bernoulli}(\overline{M}/Q)$. Because a Poisson random variable can be formed by taking an infinite sum of Bernoulli random variables (or a Binomial random variable with an infinite number of terms) with finite expectation, $[A^T D1]_n$ has the same distribution as

$$\lim_{Q\to\infty} \sum_{q=1}^{Q} w_q^{(Q)} \frac{\gamma_n(\tau_q)}{\lambda_{\bar{x}}(\tau_q)} \tag{2.26}$$

when we consider an independent copy of τ_S with Q events.

It is easily shown that

$$\mathbb{E}\left(w_q^{(Q)} \frac{\gamma_n(\tau)}{\lambda_{\bar{x}}(\tau)} \right) = \frac{\overline{M}}{Q} \int_{\mathbb{T}} \frac{\gamma_n(t)}{\lambda_{\bar{x}}(t)} f_{\tau_m}(t) dt = \frac{b_n}{Q}$$

$$\mathbb{E}\left(\left(w_q^{(Q)} \frac{\gamma_n(\tau)}{\lambda_{\bar{x}}(\tau)} \right)^2 \right) = \frac{\overline{M}}{Q} \int_{\mathbb{T}} \frac{\gamma_n^2(t)}{\lambda_{\bar{x}}^2(t)} f_{\tau_m}(t) dt \leq \frac{\Gamma_{nn}}{Q\lambda_{\min}}.$$

We use the above statements to control the expectations

$$\mathbb{E}\left(\frac{b}{Q} - w_q^{(Q)} \frac{\gamma(\tau)}{\lambda_{\bar{x}}(\tau)} \right) = 0,$$

$$\sum_{n\in U} \mathbb{E}\left(\left(\frac{b_n}{Q} - w_q^{(Q)} \frac{\gamma_n(\tau)}{\lambda_{\bar{x}}(\tau)} \right)^2 \right) \leq \frac{\text{Tr}(\Gamma_U)}{Q\lambda_{\min}}.$$

Further, we bound $\left\| \frac{b_U}{Q} - w_q^{(Q)} \frac{\gamma_U(t)}{\lambda_{\bar{x}}(t)} \right\|_2 \leq \left(\frac{\overline{M}}{Q} + 1 \right) \frac{\|\gamma_U\|_{2,\infty}}{\lambda_{\min}}$. The matrix Bernstein inequality [37] (applied to a $s \times 1$ vector) assures us that

$$\mathbb{P}\left(\left\| b_U - \sum_{q=1}^{Q} w_q^{(Q)} \frac{\gamma_U(\tau_q)}{\lambda_{\bar{x}}(\tau_q)} \right\|_2 \geq \eta \right) \leq (s+1) \exp\left(\frac{-\eta^2}{2\frac{\text{Tr}(\Gamma_U)}{\lambda_{\min}} + \frac{2}{3}\eta \left(\frac{\overline{M}}{Q} + 1 \right) \frac{\|\gamma_U\|_{2,\infty}}{\lambda_{\min}}} \right).$$

We solve for the argument of the exponent and use the concavity of the square root ($\sqrt{a^2+b^2} \leq$

$|a| + |b|$) to write

$$\mathbb{P}\left(\left\|b_U - \sum_{q=1}^{Q} w_q^{(Q)} \frac{\gamma_U(\tau_q)}{\lambda_{\bar{x}}(\tau_q)}\right\|_2 \geq \tfrac{2}{3}\left(\tfrac{\overline{M}}{Q}+1\right)\tfrac{\zeta\|\gamma_U\|_{2,\infty}}{\lambda_{\min}} + \sqrt{2\zeta\tfrac{\mathrm{Tr}(\Gamma_U)}{\lambda_{\min}}}\right) \leq (s+1)\exp(-\zeta).$$

Because $A_U^T D1$ has the same distribution as (2.26), we let $Q \to \infty$ to complete the proof.

2.12.2 Proof of Lemma 2

As a shorthand, define $G = A^T D A$ and $G_{ij}^{(m)} = \frac{\gamma_i(\tau_m)\gamma_j(\tau_m)}{\lambda_{\bar{x}}(\tau_m)}$ so that $G = \sum_{m=1}^{M} G^{(m)}$. With this definition,

$$\mathbb{E}\left(G_{ij}^{(m)}\right) = \mathbb{E}\left(\frac{\gamma_i(\tau)\gamma_j(\tau)}{\lambda_{\bar{x}}(\tau)}\right) = \frac{\Gamma_{ij}}{\overline{M}}.$$

$G^{(m)}$ is rank-1 and positive semidefinite. We can write and bound the spectral norm as

$$\left\|J^T G^{(m)} J\right\|_2 = \frac{\sum_{n\in U}\gamma_n^2(\tau_m)}{\lambda_{\bar{x}}(\tau_m)} \leq \frac{\|\gamma_U\|_{2,\infty}^2}{\lambda_{\min}}.$$

As in Section 2.12.1, we argue that G has the same distribution as $G' = \lim_{Q\to\infty}\sum_{q=1}^{Q} w_q^{(Q)} G^{(q)}$ when $w_q^{(Q)} \sim$ Bernoulli$(\overline{M}/Q)$.

Applying the matrix Chernoff inequality [37] provides the bound, for $\eta \in [0,1]$,

$$\mathbb{P}\left(\sigma(J^T G' J) \leq (1-\eta)\sigma(\Gamma_U)\right) \leq s\exp\left(-\frac{\lambda_{\min}\sigma(\Gamma_U)\eta^2}{2\|\gamma_U\|_{2,\infty}^2}\right).$$

The proof is completed by solving for the argument of the exponent and letting $Q \to \infty$ so that the distributions of G and G' converge.

CHAPTER 3

HAWKES PROCESS ESTIMATION

A Hawkes process is distinguished from other types of point processes by possessing a linear dependence on previous events. Specifically, each event has a linear effect on the future incidence of events. In the most traditional sense, the autoregression in a Hawkes process is purely excitatory, in that previous events cannot decrease the intensity at any point in the future. However, variations on the Hawkes process have been proposed to loosen this restriction and allow inhibition, e.g., [23].

Knowing the influence that each node exerts allows us to predict how likely nodes are to react to activity. To quantify these influences, we form a regression problem that fits excitation parameters to the observed activity. While this procedure is a ubiquitous, its performance and reliability is poorly understood. We present novel asymptotic recovery bounds that characterize the quality of this inference.

In an effort to improve this understanding, we undertake the task of exploring the asymptotic performance of the maximum likelihood estimator. Our analysis results in a novel asymptotic bound that focuses on how the properties of the Hawkes process impact estimation error. In particular, we attempt to explore how the structure of inter-node influence in the network impacts estimation.

3.1 Related work

In [36] it is established that the maximum likelihood estimator is consistent, asymptotically normal, and efficient for the purposes of estimating the excitation of Hawkes processes. This is a key factor in our asymptotic analysis.

Oracle inequalities providing a bound on which the intensity of a Hawkes process is estimated, using a LASSO-style estimator to estimate linear weights, are presented in [33].

This result is expanded upon in [23], which provides a similar estimation procedure that offers a high-probability bound on the error in estimating the underlying intensity. While our work is based on casting the Hawkes process as a branching process, they instead treat it as a thinning process and are thus able to consider suppressive terms that our approach cannot. Doing so requires them to introduce a nonlinear link function into the intensity function, to prevent suppression from producing a negative intensity that invalidates the point process. Both [33] and [23] differ from our result in that we instead seek to use a maximum-likelihood estimator to estimate these weights. Rather than examining the accuracy of density estimation, we instead explore the accuracy with which the excitation matrix can be estimated.

Excitation estimation is analyzed in [38]. However, they consider the separate case of *log-linear* autoregressive point processes. Like [23], they incorporate nonlinear link functions in order to ensure that the process remains well-behaved in the presence of both excitatory and suppressive interactions. Specifically, in the log-linear case a saturation effect is needed to prevent runaway excitation (leading to infinite intensity). Other works [26] further explore this space by including the possibility of learning the link function from the observations.

Although [39] pursues very different ends, namely a protocol for manipulating these networks to exhibit a desired behavior, they reach a similar conclusion to the one we achieve in Section 3.5. However, their result is, superficially, very different. We find our result much more amenable to further analysis than the one presented there.

The work of [40] explores nonasymptotic estimation of Poisson processes. While the analysis there does not apply to Hawkes processes, the similarity between the models means that a similar approach may be useful in producing a nonasymptotic version of the results we present here.

3.2 Excitation estimation for Hawkes processes

Suppose that we know $\mu_i(t)$ and the shape, but not magnitude, of $\gamma_{ij}(t)$ for every i and j. Suppose, further, that the shape of $\gamma_{ij}(t)$ is the same for all i and j so that they only differ in scale. We defer discussion of the more general case, when we wish to also estimate $\mu_i(t)$ and the shape of $\gamma_{ij}(t)$, to Section 3.8. We also discuss the general case when $\gamma_{ij}(t)$ are not all the same shape in Section 3.8.

Let the known shape of $\gamma_{ij}(t)$ be $\phi(t)$, a nonnegative and strictly causal function normalized so that $\int \phi(t)dt = 1$. In this way, we can write $\gamma_{ij}(t) = A_{ij}\phi(t)$ for an unknown excitation matrix A_{ij}, which we wish to estimate. We remark that $\phi(t)$ is a valid probability density function for a nonnegative random variable (it is nonzero on the support of positive arguments, is nonnegative, and integrates to unity).

The negative log-likelihood of a set of observations given a point process intensity $\lambda(t)$ is

$$\mathcal{L}(\tau_S|\lambda) = \int_{\mathbb{R}} \lambda(t)dt - \sum_{k\in S} \log(\lambda(\tau_k)). \tag{3.1}$$

In our case we, wish to estimate the excitation parameters A_{ij} based on intensity

$$\lambda_i(t) = \mu_i(t) + \sum_{j\in\mathbb{Z}_N} A_{ij} \sum_{k\in S_j} \phi(t-\tau_k). \tag{3.2}$$

We will minimize the negative log-likelihood of the ensemble

$$\mathcal{L}(\tau_S|\{\lambda_i\}_{i\in\mathbb{Z}_N}) = \sum_{i\in\mathbb{Z}_N} \mathcal{L}(\tau_{S_i}|\lambda_i)$$

by optimizing the A_{ij}. We are thus performing linear estimation with the basis functions $\sum_{k\in S_j} \phi(t-\tau_k)$ for each $j \in \mathbb{Z}_N$. Note that $\lambda_i(t)$ depends only on one row of A and thus the estimation of A can be accomplished independently for each row. We will define row i of A as the vector $a_i \in \mathbb{R}_+^N$.

Let us estimate the true excitation matrix A using the maximum likelihood estimate

$$\widehat{a}_i = \underset{a_i' \in \mathbb{R}_+^N}{\arg\min}\, \mathcal{L}(\tau_{S_i}|\mu_i, a_i'). \tag{3.3}$$

For a Poisson process, the accuracy with which the rows a_i are estimated is addressed in Chapter 2 or [40]. However, the assumptions there are violated by Hawkes processes because of their autoregressive behavior.

In order to characterize the performance of the maximum likelihood estimator for the parameters of Hawkes processes, we will turn our attention to the Cramér-Rao lower bound. The Cramér-Rao lower bound is based on the Fisher information matrix, which for our problem is defined to be The Fisher information matrix for the estimation of a_i is defined

$$\begin{aligned}
\mathfrak{I}_{mn}^{(i)} &= \left[\mathbb{E}\left(\nabla_{a_i}^2 \mathcal{L}(\tau_{S_i}|\mu_i, a_i)\middle|\mu_i, a_i\right)\right]_{mn} \\
&= \left[\mathbb{E}\left(\nabla_{a_i}^2 \sum_{h\in S_i} \log(\lambda_i(\tau_h|\tau_S))\middle|\mu_i, a_i\right)\right]_{mn} \\
&= \mathbb{E}\left(\sum_{h\in S_i} \frac{\left(\sum_{k\in S_m}\phi(\tau_h-\tau_k)\right)\left(\sum_{\ell\in S_n}\phi(\tau_h-\tau_\ell)\right)}{\lambda_i^2(\tau_h|\tau_S)}\right) \\
&= \mathbb{E}\left(\sum_{h\in S_i}\sum_{k\in S_m}\sum_{\ell\in S_n} \frac{\phi(\tau_h-\tau_k)\phi(\tau_h-\tau_\ell)}{\lambda_i^2(\tau_h|\tau_S)}\right).
\end{aligned} \tag{3.4}$$

The third equality follows from the derivative $\frac{d}{dx}\frac{d}{dy}\log(\lambda(x,y)) = \frac{d\lambda(x,y)}{dx}\frac{d\lambda(x,y)}{dy}\lambda^{-2}(x,y)$ and the last from the distributive property of summation. We find this matrix very difficult to analyze. Instead, we will shift our focus to the matrix

$$\Phi_{mn}^{(i)} = \mathbb{E}\left(\sum_{h\in S_i}\sum_{k\in S_m}\sum_{\ell\in S_n} \frac{\phi(\tau_h-\tau_k)\phi(\tau_h-\tau_\ell)}{\lambda_i(\tau_h|\tau_S)}\right),$$

which differs by a factor of $\lambda_i(\tau_h|\tau_S)$ in the denominator.

Because $\phi(t)$ is zero for nonpositive arguments, the inner sum is only nonzero when h corresponds to the last event among h, k, and ℓ. Defining the set $S'(t) = \{k \in S : \tau_k < t\}$,

all terms with $k \notin S'(\tau_h)$ or $\ell \notin S'(\tau_h)$ can be neglected for their lack of contribution. For the same reason, we can make use the equivalence $\lambda_i(t|\tau_S) = \lambda_i(t|\tau_{S'(t)})$. Accordingly,

$$\Phi_{mn}^{(i)} = \mathbb{E}\left(\sum_{k\in S_m}\sum_{\ell\in S_n}\mathbb{E}\left(\sum_{h\in S_i}\frac{\phi(\tau_h-\tau_k)\phi(\tau_h-\tau_\ell)}{\lambda_i(\tau_h|\tau_{S'(\tau_h)})}\middle|\tau_{S'(\tau_h)}\right)\right). \tag{3.5}$$

The density of events $h \in S_i$ is given by $\lambda_i(t|\tau_{S'(t)})$. By the definition of intensity (1.2),

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\mathbb{E}\left(\sum_{h\in S_i}\mathcal{I}_{(t\le\tau_h<t+\Delta)}\middle|\lambda_i(t,\tau_{S'(t)})\right) = \lambda_i(t,\tau_{S'(t)}),$$

from which we see that

$$\mathbb{E}\left(\sum_{h\in S_i}\frac{\phi(\tau_h-\tau_k)\phi(\tau_h-\tau_\ell)}{\lambda_i(\tau_h|\tau_{S'(\tau_h)})}\middle|\tau_{S'(\tau_h)}\right) = \int\left(\frac{\phi(t-\tau_k)\phi(t-\tau_\ell)}{\lambda_i(t|\tau_{S'(t)})}\right)\lambda_i(t|\tau_{S'(t)})dt.$$

Substituting this into (3.5), we see that

$$\Phi_{mn} = \mathbb{E}\left(\sum_{k\in S_m}\sum_{\ell\in S_n}\int\phi(t-\tau_k)\phi(t-\tau_\ell)dt\right), \tag{3.6}$$

where we have discarded the superscript from $\Phi^{(i)}$ because we now see that there is no dependence on i. We refer to this matrix as the *expected Hawkes Gram matrix* because it is the expectation of the Gram matrix of the set of (random) basis functions $\sum_{k\in S_m}\int\phi(t-\tau_k)$ for the estimation problem (3.3).

If we define $\|\lambda_i^{-1}\|_\infty^{-1} = \inf_t\lambda_i(t|\tau_S)$ and $\|\lambda_i\|_\infty = \sup_t\lambda_i(t|\tau_S)$, we note that Φ is related to $\mathfrak{I}^{(i)}$ by the semidefinite ordering

$$\Phi\|\lambda_i\|_\infty^{-1} \preceq \mathfrak{I}^{(i)} \preceq \Phi\|\lambda_i^{-1}\|_\infty. \tag{3.7}$$

The Cramér-Rao lower bound states that any unbiased estimator $\tilde{a}$ satisfies the inequal-

ity

$$\mathbb{E}\left(\|\tilde{a}_i - a_i\|_2^2\right) \geq \mathrm{Tr}((\mathfrak{I}^{(i)})^{-1}). \tag{3.8}$$

Because the maximum likelihood estimator is asymptotically efficient [36], the estimate (3.3) achieves this bound (i.e., the inequality holds with equality) as the number of observations goes to infinity. Using the relationship (3.7), in the large-sample regime it holds that

$$\mathrm{Tr}(\Phi^{-1})\|\lambda_i^{-1}\|_\infty^{-1} \leq \mathbb{E}\left(\|\widehat{a}_i - a_i\|_2^2\right) \leq \mathrm{Tr}(\Phi^{-1})\|\lambda_i\|_\infty. \tag{3.9}$$

In the purely additive case that we consider ($\inf_t \gamma_{ij}(t) = 0$), we can bound $\|\lambda_i^{-1}\|_\infty^{-1} \geq \inf_t \mu_i(t)$. It is not so trivial to bound $\|\lambda_i\|_\infty$, but under certain conditions this can be accomplished using the methods described in [33, 41]. What remains is to understand the expected Gram matrix Φ, which will be the focus of Section 3.5, and to tie this into asymptotic estimation error bounds, which we do in Section 3.6. Our analysis is based upon Galton-Watson branching processes, which we discuss in Section 3.3, and their relationship to Hawkes processes, which we discuss in Section 3.4.

3.3 Galton-Watson branching processes

Much of our analysis will draw upon the relationship between the (multivariate) Hawkes process and the (multitype) Galton-Watson branching process (GWBP). Here, we will introduce the multitype GWBP before discussing its relationship to Hawkes processes in Section 3.4. Much of the terminology for Galton-Watson processes is related to that of family trees, as they were originally developed to model the survival and extinction of surnames [42].

In a multitype GWBP [43], we consider a population with different *types* of member. This population evolves over a series of generations. At each generation, every member produces offspring for the next generation. The statistics of this reproduction are based only on the member's type. Generation occurs independently for each member, so this

is a Markov process, and a member need-not produce offspring of only its same type. This process runs until *extinction*, when there are no members in the current generation to produce the next. It is possible that extinction never occurs, in which case the process continues indefinitely.

Let us define the length-N population vector z_t, which counts the number of members of each of N types in generation t ($[z_t]_i$ being the the number of type-i members). We will usually assume that the population of the first generation z_0 is given explicitly. The sequence $\{z_t\}_{t=0}^{\infty}$ is the realization of a GWBP, which we will refer to as a *branch*.

The most significant statistic relating to the asymptotic behavior of the process is the expectation $\mathbb{E}(z_{t+1}|z_t)$. A property of a GWBP (resulting from the fact that reproduction is independent and identical for each member of each type) is that $\mathbb{E}(z_{t+1}|z_t) = Az_t$ for a $A \in \mathbb{R}_+^{N\times N}$. We will call the matrix A an excitation matrix, as it serves a virtually identical role the the excitation matrix in a Hawkes processes. We will elaborate on this relationship in Section 3.4. If we define the *branch population* to be $z_\infty = \sum_{t=0}^{\infty} z_t$ and the *branch excitation matrix* $B = \sum_{t=0}^{\infty} A^t$ (noting that $B \in \mathbb{R}_+^{N\times N}$), then iterated expectation reveals that $\mathbb{E}(z_\infty|z_0) = Bz_0$.

Note that B is only finite if $\rho(A) < 1$, in which case $B = (I - A)^{-1}$ and $AB = BA = B - I$. In this regime, we say that the GWBP is *subcritical*. For subcritical GWBPs, it holds that $\lim_{t\to\infty} \mathbb{P}(z_t = 0) = 1$ for any z_0. For a *critical* process with $\rho(A) = 1$ or a *supercritical* process with $\rho(A) > 1$, there exist z_0 for which $\mathbb{E}(z_\infty|z_0) = \infty$. In words, a subcritical process will almost-surely become extinct while a (super)critical process has a finite chance of producing an infinite number of events (at least under some initial conditions). As such, we cannot reliably expect to observe a (super)critical process in its entirety.

The GWBP we will find relevant to Hawkes processes is one with a Poisson offspring distribution. Let the number of offspring of type i produced by a type-j member of the previous generation be given by a random variable with distribution Poisson(A_{ij}). In this

case, the intergenerational distribution is

$$z_{t+1} \sim \text{Poisson}(Az_t). \tag{3.10}$$

Assuming this distribution, one can show that the next-generation inter-type correlation matrix is

$$\mathbb{E}(z_{t+1}z_{t+1}^\top | z_t) = Az_t z_t^\top A^\top + \text{diag}(Az_t), \tag{3.11}$$

which is computed by simply considering the statistics of a vector of independent Poisson random variables.

One will recognize the apparent similarity between (1.13), corresponding to the LARPP, and (3.10). In Section 3.4, we will show that this similarity also extends to general Hawkes processes. Furthermore, we will show that a GWBP characterizes significant aspects of the behavior of a Hawkes process.

3.4 Branching interpretation of Hawkes processes

The autoregressive excitation of Hawkes processes results in observations (events) that tend to cluster in t. The structure of these clusters is most succinctly analyzed using the branching representation of the process [41, 44].

3.4.1 Poisson decomposition

For any point process, we can use the Poisson superposition property to decompose a point process into multiple constituent processes [20]. If a point process has intensity $\lambda(t) = \lambda_1(t) + \lambda_2(t) + \ldots$, this property provides that a realization of (i.e., the set of events generated by) this process is statistically equivalent to the union of the realizations of point processes with respective intensities $\lambda_1(t)$, $\lambda_2(t)$, ...

Thus, we can decompose a Hawkes process into many constituent point processes [21, 41, 44]. Instead of considering the single intensity function $\lambda_i(t|\tau_S)$ from (1.12), we can

treat it as a union of many constituent processes with intensities $\mu_i(t)$ and $\gamma_{ij}(t-\tau_k)$ for all $j \in \mathbb{Z}_N$ and $k \in S_j$. Given an event $k \in S_j$, $\gamma_{ij}(t-\tau_k)$ is contributed to the intensity of subprocess i. Because this contribution is fixed (given τ_k), it contributes as a Poisson process. Although the events, themselves, can contribute to the generation of additional events, this decomposition allows us to treat a Hawkes process as a cascade of Poisson processes. The cascading aspect is modeled by a GWBP.

3.4.2 Generating Hawkes processes

Suppose that we generate a Poisson process using each base intensity $\mu_i(t)$. These processes have absolutely no dependence on anything but the predefined base intensities. As a Poisson processes, they also have no dependence on themselves. We will refer to events derived from the $\mu_i(t)$ constituents as *innovations*.

Suppose that each innovation gives birth to N new Poisson processes. If the event occurs at coordinate τ_k with $k \in S_j$, then there is one new Poisson process spawned for each $i \in \mathbb{Z}_N$ and these have intensities $\gamma_{ij}(t-\tau_k)$. Based on our definition of the excitation matrix A, a type-j event will result in Poisson(A_{ij}) new type-i *excitation* events. If the parent is located at coordinate τ_k then the distribution of these excitation events is given by the probability density function $f_t(t|\tau_k) = A_{ij}^{-1}\gamma_{ij}(t-\tau_k)$. Each excitation event is allowed to reproduce in exactly the same way that innovations reproduce.

The collection of all innovation and excitation events generated by this branching algorithm will form a Hawkes process. This algorithm is presented in pseudocode in Algorithm 1. In this construction, each constituent processes is allowed to create additional constituent processes. One will recognize that the reproduction structure described here is exactly that of the Poisson GWBP described in Section 3.3. The only difference is that we have, additionally, assigned each member a coordinate τ_k. Accordingly, we now have cause to track individual events rather than just the per-generation population as we described in the GWBP. Note that it is also possible to generate realizations of Hawkes processes based

Algorithm 1 Hawkes process generation via branchin

initialize: $\tau = \varnothing$, $S_1 = S_2 = \dots S_N = \varnothing$
for $i \in \mathbb{Z}_N$ **do**
  call: *BRANCH*$(i, \mu_i(t))$ {generate type-i innovations}
return $\tau, S_1, S_2, \dots S_N$

subroutine: *BRANCH*$(j, \lambda(t))$ {generate type-j events from intensity $\lambda(t)$}
$(\phi_{S'}, S') \sim \text{PoissonProc}(\lambda(t), \mathbb{R})$ {generate via Poisson process}
$\tau = \tau \cup \phi_{S'}$, $S_j = S_j \cup S'$ {add to record}
for $k \in S'$ **do**
  for $i \in \mathbb{Z}_N$ **do**
    call: *BRANCH*$(i, \gamma_{ij}(t - \phi_k))$ {generate type-i offspring of event k}
return

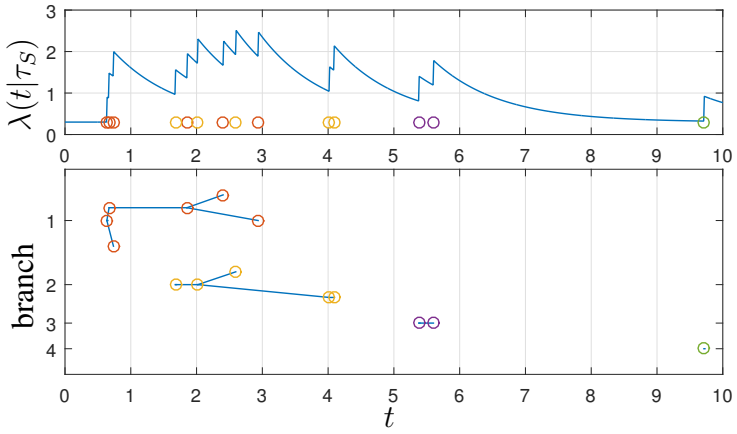

Figure 3.1: Intensity function (top), branch structure (bottom), and events (both) of a Hawkes process realization with four branches

on other mechanisms such as, for example, rejection sampling as in Section 1.3.7.

We refer to an innovation and all excitations that are direct or indirect descendants of it as a *branch*. We present an example of a (univariate) Hawkes process and its underlying branch structure in Figure 3.1. Like in a GWBP, if each event produces (on average) more than one additional event then branches may never terminate. In this paper, we exclude such a regime by requiring Hawkes processes to be subcritical (i.e., satisfy $\rho(A) < 1$).

3.4.3 Notes on the branching interpretation

A branch can be thought of as a family tree, and much of our terminology will reflect this. We will use the term *children* to refer to the events directly descended from a *parent* and

descendants more generally to refer to the children, children-of-children, etc., of an event. We use the term *ancestor* to represent a parent, parent-of-parent, etc. We will also usually consider an event to be both an ancestor and descendant of itself. We say that events *cause* their descendants. We say that the *last common ancestor* of an event pair is the last event that is an ancestor of both. An event is the last common ancestor of itself paired with any of its descendants (or itself). We say that two events are *cousins* under a third if the third is the last common ancestor of the two. By this slightly-unusual definition, an event and its ancestor are cousins under the ancestor.

It is important to note that it is typically impossible to determine the branch structure when observing a Hawkes process. For example, the events of branches 1, 2, and 3 in Figure 3.1 could convincingly be called members of the same branch (because the events appear to form a single cluster) or could otherwise have their members shuffled. The underlying branch structure of a Hawkes process can be thought of as a set of hidden state variables. Although it is irrelevant to our analysis, it is possible to estimate the *distribution* of these state variables (e.g., the *probability* that one event is the parent of another) under this interpretation, as discussed and explored in [45] and Chapter 5.

Despite the fact that we do not usually observe the branch structure of Hawkes processes, the branching interpretation is extremely useful for statistical analysis. While the Hawkes process itself is highly self-dependent, the branching interpretation allows us to consider a number of cases where events are independent or conditionally independent of each other. Innovations are independent of each other, as they arise from the Poisson processes described by the $\mu_i(t)$. Conditioned on an event, that event's children are independent of each other. From these two facts we derive two statements that will be key to determining the basic statistics of Hawkes processes. First, any two events from different branches (i.e., events that result from, or are, different innovations) are independent of each other. Second, within a branch, any two events are independent when conditioned on their last common ancestor.

3.4.4 Branch intensity function

In Section 3.4.2, we described how events "reproduce" in a Hawkes process. To this end, it will be important for us to understand the "average" reproduction patterns of the process. Specifically, we will characterize the behavior of the process when conditioning on one or more initial events. This will serve an analogous role to the branch intensity matrix for a GWBP.

Recall that the impact of a type-j event at τ is to increase the intensity function of the subprocess i by $\gamma_{ij}(t-\tau)$. Thus, if we are given some initial set of events, the density of events in the next generation is given by the current density convolved with the $\gamma_{ij}(t)$ functions. Suppose that our initial set of events is a single type-j event at coordinate $t=0$. Let $r_{ij}^{(k)}(t)$ be the density of type-i events in generation k given a single type-j event at $t=0$, $k=0$. Our initial event corresponds to $r_{ij}^{(0)}(t) = \mathcal{I}_{(i=j)}\delta(t)$. The density of later generations, conditioned on this original event, are given by the recurrence

$$r_{ij}^{(k)}(t) = \sum_{\ell \in \mathbb{Z}_N} r_{\ell j}^{(k-1)}(t) * \gamma_{i\ell}(t) = \sum_{\ell \in \mathbb{Z}_N} \int r_{\ell j}^{(k-1)}(\tau)\gamma_{i\ell}(t-\tau)d\tau.$$

The *branch intensity function* is computed by combining the conditional densities of all generations

$$\beta_i^{(j)}(t) = \sum_{k=0}^{\infty} r_{ij}^{(k)}(t) \tag{3.12}$$

and represents the density of type-i events in a branch (i.e., all generations) when conditioned on a type-j ancestor at $t=0$. In this situation, $r_{ij}^{(k)}(t)$ is analogous to the GWBP quantity $[A^k]_{ij}$ and

$$\int \beta_i^{(j)}(t)dt = \sum_{k=0}^{\infty} [A^k]_{ij} = B_{ij}. \tag{3.13}$$

Thus, as we saw when discussing GWBPs, B represents the expected number of type-i events resulting (over multiple generations) from an event of type j.

3.4.5 A priori intensity

Equipped with the branch intensity function, we can also consider the average behavior of a Hawkes process. Recall that type-i innovations have density $\mu_i(t)$. Because the branch intensity function gives the expected response to an initial event and this density describes a set of initial events, the *a priori intensity function*

$$\eta_i(t) = \sum_{j \in \mathbb{Z}_N} \mu_j(t) * \beta_i^{(j)}(t) = \sum_{j \in \mathbb{Z}_N} \int \mu_j(\tau) \beta_i^{(j)}(t-\tau) d\tau, \tag{3.14}$$

represents the density of subprocess i before observing any events. In other words, the a priori intensity satisfies $\eta_i(t) = \mathbb{E}(\lambda_i(t))$. Recalling our definition of the total base intensity, $u_i = \int \mu_i(t) dt$, it follows that

$$\mathbb{E}\left(\int \lambda_i(t) dt \right) = \int \eta_i(t) dt = [Bu]_i, \tag{3.15}$$

which is the a priori expected number of type-i events.

3.5 Expectation of the Hawkes Gram matrix

In this section, we present and derive the expression for computing the expected Gram matrix Φ. As we discussed in Section 3.2, the matrix Φ is of central importance in understanding the the accuracy with which the excitation matrix A is estimated.

Define the cross-correlation functions $R_\phi(\tau) = \int \phi(t)\phi(t-\tau)dt$, $R_{\eta mn}(\tau) = \int \eta_m(t)\eta_n(t-\tau)dt$, $R^{(j)}_{\beta mn}(\tau) = \int \beta^{(j)}_m(t)\beta^{(j)}_n(t-\tau)dt$. With these definitions, we present one of our main results:

Theorem 3 *Consider a Hawkes process with intensity given by* (3.2) *where $\mu_i(t)$, A_{ij}, and $\phi(t)$ are nonnegative. When we define $u_i = \int \mu_i(t)dt$, the expected Hawkes Gram matrix*

has entries

$$\Phi_{mn} = \int R_\phi(\tau) \left(R_{\eta mn}(\tau) + \sum_{j \in \mathbb{Z}_N} [(I - A)^{-1} u]_j R^{(j)}_{\beta mn}(\tau) \right) d\tau. \tag{3.16}$$

We devote Section 3.5.1 to the proof of this theorem and offer some computational remarks in Section 3.5.2.

A comparable result is derived via a different method in [39], yielding a substantially different form. We will find the form of (3.16) to be much more interpretable when we discuss estimation performance in Section 3.6.

3.5.1 Proof of Theorem 3

Partitioning event pairs

We remark that it is possible to rewrite the expression for the Gram matrix (3.6) as

$$\Phi_{mn} = \mathbb{E} \left(\sum_{k \in S_m} \sum_{\ell \in S_n} R_\phi(\tau_k - \tau_\ell) \right). \tag{3.17}$$

Computing this matrix will require us to understand the distribution of $\tau_k - \tau_\ell$ for all pairs of events. As we already discussed, the challenge lies in the interdependence of many (τ_k, τ_ℓ) pairs.

Define the set of event pairs where the events are from the same branch to be $\mathcal{Q}$, so that $(k, \ell) \in \mathcal{Q}$ if and only if events k and ℓ are from the same branch. Also, define $\mathcal{S}_{mn} = \{(k, \ell) : k \in S_m, \ell \in S_n\}$. We will find it helpful to partition the expectation into interbranch ($(k, \ell) \notin \mathcal{Q}$) and intrabranch ($(k, \ell) \in \mathcal{Q}$) pairs,

$$\Phi_{mn} = \mathbb{E} \left(\sum_{(k,\ell) \in \mathcal{S}_{mn} \setminus \mathcal{Q}} R_\phi(\tau_k - \tau_\ell) \right) + \mathbb{E} \left(\sum_{(k,\ell) \in \mathcal{S}_{mn} \cap \mathcal{Q}} R_\phi(\tau_k - \tau_\ell) \right), \tag{3.18}$$

and address each part separately. Above, $\mathcal{S}_{mn} \setminus \mathcal{Q}$ denotes the intersection of $\mathcal{S}_{mn}$ and the

complement of $\mathcal{Q}$.

We discuss the terms in the following subsections.

Interbranch event pairs

Here we will compute the first term of (3.18), corresponding to interbranch event pairs. As discussed in Section 3.4, two events that belong to different branches are statistically independent. In other words, τ_k and τ_ℓ are independent if $(k, \ell) \notin \mathcal{Q}$. Because the events are independent, τ_k and τ_ℓ are distributed proportionally to their respective prior intensities $\eta_m(t)$ and $\eta_n(t)$. The distribution of $\tau_k - \tau_\ell$ can thus be computed by simply convolving the two probability densities (normalized intensities) after reversing the one corresponding to ℓ (because τ_ℓ is substracted), yielding the probability density

$$f_{\tau_k - \tau_\ell}(\tau | (k, \ell) \in \mathcal{S}_{mn} \setminus \mathcal{Q}) = \frac{\int \eta_m(t)\eta_n(t - \tau)dt}{[Bu]_m[Bu]_n}. \tag{3.19}$$

Knowing the distribution, now we only need to consider the expected number of pairs in the set $\mathcal{S}_{mn} \setminus \mathcal{Q}$. For this, we need only consider GWBPs. Suppose that we define $\bar{u} = \sum_{j \in \mathbb{Z}_N} u_j$ (the expected total number of innovations of all types) and assign z_0 randomly according to $\mathbb{P}(z_0 = e_j) = u_j/\bar{u}$. This distribution of z_0 matches the probability that any given innovation is of a particular type. It follows that $\mathbb{E}(z_\infty) = \frac{1}{\bar{u}} Bu$. Further, if we consider the populations of L independent and identically distributed branch populations $z_\infty^{(\ell)}$ for $\ell \in \mathbb{Z}_L$ then

$$\mathbb{E}\left(\sum_{k \in \mathbb{Z}_L} \sum_{\ell \neq k} z_\infty^{(k)} z_\infty^{(\ell)\top} \middle| L \right) = \frac{L(L-1)}{\bar{u}^2} Buu^\top B^\top.$$

By Poisson superposition, the total number of innovations (and thus branches) in our pro-

cess is distributed as $L \sim \text{Poisson}(\bar{u})$, which yields

$$\mathbb{E}\left(\sum_{k\in\mathbb{Z}_L}\sum_{\ell\neq k} z_\infty^{(k)} z_\infty^{(\ell)\top}\right) = Buu^\top B^\top.$$

From this, we conclude that the expected number of event pairs in the set $\mathcal{S}_{mn} \setminus \mathcal{Q}$ is $[Bu]_m[Bu]_n$ and

$$\mathbb{E}\left(\sum_{(k,\ell)\in\mathcal{S}_{mn}\setminus\mathcal{Q}} R_\phi(\tau_k - \tau_\ell)\right) = \int R_\phi(\tau)\int \eta_m(t)\eta_n(t-\tau)dtd\tau. \tag{3.20}$$

Intrabranch event pairs

Now we will continue by addressing the second term of (3.18), corresponding to the intrabranch event pairs. In Section 3.4, we argued that two events from the same branch are conditionally independent given their last common ancestor. Note that any two events on a branch have a last common ancestor. In the case that we are considering one event and either itself or one of its descendants, we say that it is the last common ancestor of the pair. Define the set $\mathcal{P}_q$ such that $(k,\ell) \in \mathcal{P}_q$ if event q is the last common ancestor of k and ℓ. Note that the set of $\{\mathcal{P}_q\}_{q\in S}$ forms a partition of $\mathcal{Q}$ (every event pair in $\mathcal{Q}$ belongs to exactly one $\mathcal{P}_q$) so that

$$\sum_{(k,\ell)\in\mathcal{S}_{mn}\cap\mathcal{Q}} R_\phi(\tau_k - \tau_\ell) = \sum_{j\in\mathbb{Z}_N}\sum_{q\in S_j}\sum_{(k,\ell)\in\mathcal{S}_{mn}\cap\mathcal{P}_q} R_\phi(\tau_k - \tau_\ell). \tag{3.21}$$

Because the coordinates of all event pairs in $\mathcal{P}_q$ are conditionally independent (given τ_q), we will find this partitioning to be extremely convenient. We can condition on an event and consider the contribution of all pairs for which it is the last common ancestor, as in

$$\mathbb{E}\left(\sum_{(k,\ell)\in\mathcal{S}_{mn}\cap\mathcal{P}_q} R_\phi(\tau_k - \tau_\ell)\middle| q \in S_j\right). \tag{3.22}$$

Because $(k, \ell) \in \mathcal{P}_q$, we have the distributions

$$f_{\tau_k - \tau_q}(\tau | (k, q) \in \mathcal{S}_{mj} \cap \mathcal{P}_q, q \in S_j) = \frac{\beta_m^{(j)}(\tau)}{B_{mj}}$$

$$f_{\tau_\ell - \tau_q}(\tau | (\ell, q) \in \mathcal{S}_{nj} \cap \mathcal{P}_q, q \in S_j) = \frac{\beta_n^{(j)}(\tau)}{B_{nj}}.$$

Because the two distributions are independent when conditioned on τ_q and $\tau_k - \tau_\ell = (\tau_k - \tau_q) - (\tau_\ell - \tau_q)$,

$$f_{\tau_k - \tau_\ell}(\tau | (k, \ell) \in \mathcal{S}_{mn} \cap \mathcal{P}_q, q \in S_j) = \frac{\int \beta_m^{(j)}(t) \beta_m^{(j)}(t - \tau) dt}{B_{mj} B_{nj}} \tag{3.23}$$

What remains is to determine the expected number of (k, ℓ) pairs in the set $\mathcal{S}_{mn} \cap \mathcal{P}_q$ when $q \in S_j$.

First, we will consider the number of pairs in the set $(k, \ell) \in \mathcal{S}_{mn} \cap \mathcal{P}_q$ when $q \in S_j$ and $k = q$ or $\ell = q$. If $k = \ell = q$ (which requires $m = n = j$), there is one such pair. If $k = q$ (which requires $m = j$), there are B_{nj} pairs in expectation. If $\ell = q$ (which requires $n = j$), there are B_{mj} pairs in expectation. Thus, when at-least one of the events is q, there are $\mathcal{I}_{(m=j)} B_{nj} + \mathcal{I}_{(n=j)} B_{mj} - \mathcal{I}_{(m=n=j)}$ expected pairs.

Second, we will consider the number of pairs in the set $\mathcal{S}_{mn} \cap \mathcal{P}_q$ when $q \in S_j$ and $k \neq q$ and $\ell \neq q$. Because neither event is the last common ancestor, both must be descended from (or be) a different child of q. The expected number of first generation child pairs (by type, according to their position in the matrix) is given by (3.11) to be $A e_j e_j^\top A^\top + \mathrm{diag}(A e_j)$. Removing the pairs that count the same child twice (of which there are Poisson$(A e_j)$ on the diagonal elements only), we are left with correlation matrix $A e_j e_j^\top A^\top$. A type-v first-generation child has B_{mv} descendants (counting itself) of type m, in expectation. Thus, the expected number of pairs under q that do not include q is $\sum_{v,w \in \mathbb{Z}_N} A_{vj} A_{wj} B_{mv} B_{nw} = [BA]_{mj} [BA]_{nj}$. Using the observation that $AB = BA = B - I$ allows us to state that the

expected number of such pairs is $(B_{mj} - \mathcal{I}_{(m=j)})(B_{nj} - \mathcal{I}_{(n=j)})$.

Combining these two cases yields the result that the expected number of terms is

$$\mathbb{E}\left(|(k,\ell) \in \mathcal{S}_{mn} \cap \mathcal{P}_q || q \in S_j\right) = B_{mj} B_{nj}.$$

Recalling that the expected number of type-j events is given by $\mathbb{E}(|S_j|) = [Bu]_j$, we conclude that

$$\begin{aligned}\mathbb{E}\left(\sum_{(k,\ell)\in\mathcal{S}_{mn}\cap\mathcal{Q}} R_\phi(\tau_k - \tau_\ell)\right) &= \sum_{j\in\mathbb{Z}_N} \mathbb{E}\left(\sum_{q\in S_j} \mathbb{E}\left(\sum_{(k,\ell)\in\mathcal{S}_{mn}\cap\mathcal{P}_q} R_\phi(\tau_k - \tau_\ell)\middle| q \in S_j\right)\right) \\ &= \sum_{j\in\mathbb{Z}_N} [Bu]_j \int R_\phi(\tau) \int \beta_m^{(j)}(t)\beta_n^{(j)}(t-\tau) dt d\tau.\end{aligned} \tag{3.24}$$

Substituting (3.20) and (3.24) into (3.18) completes the proof of Theorem 3.

3.5.2 Computational notes

While the expression for equation (3.16) in Theorem 3 is relatively simple, it is composed of a substantial number of convolution operations. While cumbersome to compute in their native domain, convolutions are easily evaluated in the Fourier-transform domain. Here, we will provide a set of useful identities for these computations.

Let $\mathcal{F}\{\xi\}(\omega)$ represent the Fourier transform of a function $\xi(t)$ at frequency ω and let $\mathcal{F}^*\{\xi\}(\omega)$ be its complex conjugate. Using the property that $\mathcal{F}\{a * b\}(\omega) = \mathcal{F}\{a\}(\omega)\mathcal{F}\{b\}(\omega)$,

it holds that

$$\begin{aligned}
\mathcal{F}\{\beta_i^{(j)}\}(\omega) &= \sum_{k=0}^{\infty} \left[(A\mathcal{F}\{\phi\}(\omega))^k\right]_{ij} = \left[(I - A\mathcal{F}\{\phi\}(\omega))^{-1}\right]_{ij} \\
\mathcal{F}\{\eta_i\}(\omega) &= \sum_{j\in\mathbb{Z}_N} \mathcal{F}\{\mu_j\}(\omega)\mathcal{F}\{\beta_i^{(j)}\}(\omega) \\
\mathcal{F}\{R_\phi\}(\omega) &= |\mathcal{F}\{\phi\}(\omega)|^2 \\
\mathcal{F}\{R_{\eta mn}\}(\omega) &= \mathcal{F}\{\eta_m\}(\omega)\mathcal{F}^*\{\eta_n\}(\omega) \\
\mathcal{F}\{R_{\beta mn}^{(j)}\}(\omega) &= \mathcal{F}\{\beta_m^{(j)}\}(\omega)\mathcal{F}^*\{\beta_n^{(j)}\}(\omega).
\end{aligned}$$

Parseval's theorem permits us to compute (3.16) in the Fourier domain, yielding

$$\Phi_{mn} = \int \mathcal{F}^*\{R_\phi\}(\omega) \left(\mathcal{F}\{R_{\eta mn}\}(\omega) + \sum_{j\in\mathbb{Z}_N} [(I - A)^{-1}u]_j \mathcal{F}\{R_{\beta mn}^{(j)}\}(\omega)\right) d\omega. \quad (3.25)$$

Computationally, this is considerably easier than carrying out the extensive convolutions in the time domain, as was originally presented. The savings in computing $\beta_i^{(j)}(t)$ are particularly noteworthy, as the convergent sum can be computed in closed-form as a geometric series.

As an example, the univariate ($N = 1$) Hawkes process satisfies

$$\Phi = \int \frac{|\mathcal{F}\{\phi\}(\omega)|^2}{|1 - A\mathcal{F}\{\phi\}(\omega)|^2} \left(|\mathcal{F}\{\mu\}(\omega)|^2 + \frac{u}{1 - A}\right) d\omega.$$

3.6 Asymptotic estimation characterization

With the expectation provided by Theorem 3, we can make asymptotic estimation guarantees for the excitation matrix A of a Hawkes process. While the algebra is sufficiently complicated to make general statements tedious, here we will present some instructive examples. While the analysis of Section 3.5 supports arbitrary base intensity functions, we will, for brevity, limit our discussion to Hawkes processes with constant base intensities.

3.6.1 Stationary process

If we restrict our base intensity to be a rectangle function $\mu_i(t) = \mathcal{I}_{(0\leq t\leq T)}v_n$ (noting that $u = Tv$) then we can make the approximation

$$R_{\eta mn}(\tau) \approx \max\{T - |\tau|, 0\}[Bvv^\top B^\top]_{mn}.$$

This approximation becomes arbitrarily tight as $T \to \infty$. We also assume that $\phi(t)$ is of approximately-finite extent, i.e., for any $\epsilon > 0$ there exists a finite t' such that $\int_{t'}^{\infty} \phi_{mn}(t)dt < \epsilon$. As such, $R_\phi(t)$ is (approximately) compactly supported and so, for large T, $\int R_\phi(\tau)R_{\eta mn}(\tau)d\tau \approx R_{\eta mn}(0)$. Under these assumptions, Theorem 3 tells us that

$$\lim_{T\to\infty} \frac{\Phi}{T} = Bvv^\top B^\top + \Upsilon, \tag{3.26}$$

where we define $\Upsilon_{mn} = \int R_\phi(\tau) \sum_{j\in\mathbb{Z}_N} [Bv]_j R^{(j)}_{\beta mn}(\tau)d\tau$.

Using the Sherman-Morrison matrix inversion lemma,

$$\lim_{T\to\infty} T \operatorname{Tr}\left(\Phi^{-1}\right) = \operatorname{Tr}(\Upsilon^{-1}) - \frac{v^\top B^\top \Upsilon^{-2} Bv}{1 + v^\top B^\top \Upsilon^{-1} Bv},$$

which is relevant to estimation via equation (3.9). While this value can be computed, it can also be bounded

$$\operatorname{Tr}(\Upsilon^{-1}) - \|\Upsilon^{-1}\|_2 \leq \lim_{T\to\infty} T \operatorname{Tr}\left(\Phi^{-1}\right) \leq \operatorname{Tr}(\Upsilon^{-1}). \tag{3.27}$$

3.6.2 Example: first-order LARPP

We can make slightly more concrete statements if we set an excitation kernel. For example, the first-order LARPP uses $\phi(t) = \delta(t-1)$. Because we are considering discrete coordinates, we will assume that the base intensity is a discrete boxcar function $\mu_j(t) = v_j \sum_{k=1}^{T} \delta(t-k)$. This kernel yields $\beta_i^{(j)}(t) = \sum_{k=0}^{\infty} [A^t e_j]_i \delta(t-k)$ for inte-

ger nonnegative t, $R_\phi(t) = \delta(t)$, and $R^{(j)}_{\beta mn}(0) = \sum_{t=0}^{\infty}[A^t e_j (A^t e_j)^\top]_{mn}$.

Under these definitions, $\Upsilon_{mn} = \frac{1}{T}\sum_{j\in\mathbb{Z}_N}[Bu]_j R^{(j)}_{\beta mn}(0) = \sum_{t=0}^{\infty}[A^t \operatorname{diag}(Bv)(A^t)^\top]_{mn}$. This satisfies the equality $\Upsilon = \operatorname{diag}(Bv) + A\Upsilon A^\top$, which has the solution $\operatorname{vec}(\Upsilon) = (I - A\otimes A)^{-1}\operatorname{vec}(\operatorname{diag}(Bv))$. Noting that $\operatorname{diag}(Bv) \preceq \Upsilon \preceq \frac{\|Bv\|_\infty}{1-\|A\|_2^2} I$ (which becomes tight as $A \to 0$), we can use (3.9) to bound

$$N\frac{1-\|A\|_2^2}{\|B\|_\infty} \le N\frac{1-\|A\|_2^2}{\|Bv\|_\infty}\|v\|_1 \le \lim_{T\to\infty} T\mathbb{E}\left(\|\widehat{A} - A\|_F^2\right) \le \operatorname{Tr}(\operatorname{diag}(Bv)^{-1})\|\lambda\|_{\infty,1}$$

where $\|\lambda\|_{\infty,1} = \sum_{i\in\mathbb{Z}_N}\sup_t \lambda_i(t)$. To make the lower bound of this statement, we have committed a minor violation of one of our conditions, namely that we observe the entire process. Instead, we will assume that we stop observing at time T (so that our minimum intensity vector is given by v) and that the remaining tail would have a negligible effect on the estimate (which will be true as $T\to\infty$, since the tail is almost-surely finite). The much more significant issue is that $\lim_{T\to\infty}\|\lambda\|_{\infty,1} = \infty$, despite the fact that $\lambda(t)$ reaches very large values extremely rarely [41].

If we instead conjecture that the true expectation can be bounded by something related to the mean intensity, rather than extreme intensities (by analogy to [40] this outcome appears to be plausible), it would be possible to make a statement to the effect of

$$c_1 N(1-\|A\|_2^2) \le \lim_{T\to\infty} T\mathbb{E}\left(\|\widehat{A} - A\|_F^2\right) \le c_2 \operatorname{Tr}(\operatorname{diag}(Bv)^{-1})\operatorname{Tr}(\operatorname{diag}(Bv))$$

for factors c_1 and c_2 that would depend only mildly on the system parameters.

To summarize, there is strong evidence that estimation of first-order LARPPs with long rectangular base intensities behaves like

$$\mathbb{E}\left(\|\widehat{A} - A\|_F^2\right) \lesssim \frac{\operatorname{Tr}(\operatorname{diag}(Bv)^{-1})\operatorname{Tr}(\operatorname{diag}(Bv))}{T}. \tag{3.28}$$

We note that Bv corresponds to the vector of average rates (by subprocess). As such, the

driving force in the estimation error is the ratio of the arithmetic and harmonic means of the number of events in each subprocess. Performance is improved when the number of events is balanced across types. In cases of imbalance, it appears that excitation parameters corresponding to underrepresented subprocesses will (unsurprisingly) dominate the estimation error.

3.6.3 Degeneracy

Estimation from fewer samples will yield less-accurate estimates of the excitation matrix. However, it is interesting to know what system configurations are ill-posed in such a way as to yield poor estimates even with many samples.

Equation (3.8) states that estimation is difficult when the Fisher information $\mathfrak{I}^{(i)}$ has one or more small eigenvalues. Thanks to (3.7), we can equivalently say that it is difficult when the expected Gram matrix Φ has small eigenvalues. The interesting consequence of (3.26) is that it is difficult to actually achieve a Φ with this status, even with seemingly-poor parameter configurations. Although we will only elaborate on first-order LARPPs, a similar consequence would hold in many Hawkes systems.[1]

By a poor parameter configuration, we mean that the excitation matrix and base intensity vector are chosen in such a way as to cause maximum confusion when estimating parameters. Disambiguation is difficult if, for example, multiple subprocesses behave very similarly. When attempting to infer their effect on a given subprocess, it will be difficult to determine which subprocess is responsible for what excitation. This would be reflected as small eigenvalues in the Gram matrix.

[1]The discrete-coordinate nature of LARPPs leads to low variety in event times, so they may already represent an unusually difficult case. We anticipate that other examples of Hawkes processes might generally be more robust than this example.

For example, consider the system with parameters

$$A = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, v = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Subprocesses 2 and 3 are both Poisson processes with the same constant intensity, and that intensity is somewhat high. We expect them to be similar, given that they are statistically identical. The excitation of subprocess 1 is defined by the first row of A, $a_1 = [0, 0.5, 0.5]$. Because subprocesses 2 and 3 behave identically, we could expect subprocess 1 to act similarly for any combination of influence from the two. One would expect that the choice $a_1 = [0, 1, 0]$ or $a_1 = [0, 0, 1]$ would give subprocess 1 similar statistics. Indeed, for any of the discussed choices of a_1, the average intensity of each subprocesses (indexed by i) is $[Bv]_i = 1$. The key consideration is to determine whether the random variation within the process provides enough deviation from this mean to allow the correct coefficients to be discerned.

Using (3.26), with increasing T the trace of the inverse Gram matrix for this system will converge to $\mathrm{Tr}(\Phi^{-1}) \to \frac{2.00}{T}$. If we set $A = 0$ and compensate v to maintain $[Bv]_i = 1$, we see $\mathrm{Tr}(\Phi^{-1}) \to \frac{2.25}{T}$. If we set $A = 0.9I$ and still hold Bv constant, $\mathrm{Tr}(\Phi^{-1}) \to \frac{0.501}{T}$. Note that, while the previous system appeared to be very poorly conditioned, these alternative systems with radically different structures (all subprocesses are independent) produce asymptotic errors that differ only by a modest constant factor. It appears that the raw number of events observed is a much more significant contributing factor that rapidly overcomes most challenging architectures. The conclusion is that the randomness inherent to process is typically sufficient to disambiguate excitation.

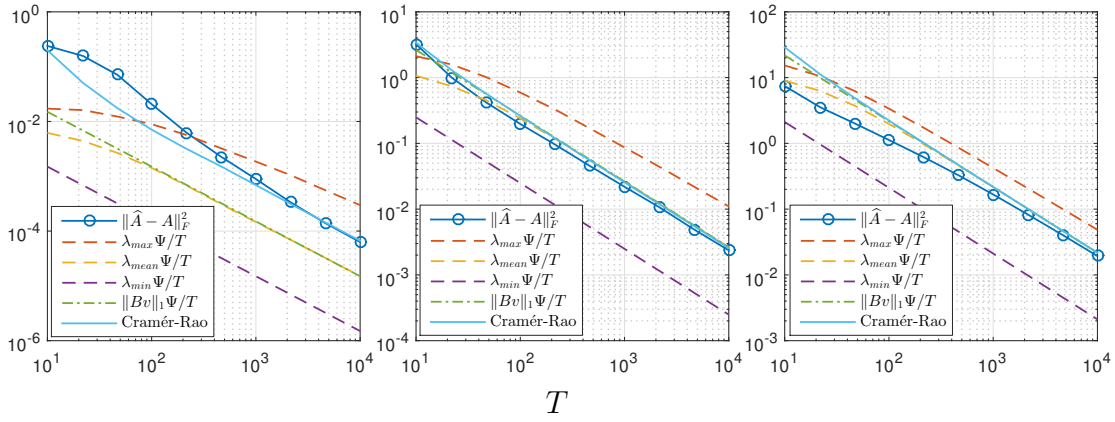

Figure 3.2: Examples comparing the number of observations to recovery error, the expected Gram matrix, and the Cramér-Rao bound for Hawkes processes with $N = 1$, $N = 5$, $N = 15$ (left to right) and defining $\Psi = \operatorname{Tr}\left((Bvv^\top B^\top + \Upsilon)^{-1}\right)$

3.7 Simulations

3.7.1 Asymptotic convergence

We have provided an asymptotic result describing the behavior as the number of observations goes to infinity. However, we have not explored the rate of convergence to this asymptotic result. To do this, we perform simulations for LARPPs with $N \in \{1, 5, 15\}$. In each case, we choose a matrix A by scaling a matrix of independent and identically distributed exponential random variables and scaling the result so that $\rho(A) = 0.9$ and we choose v as a vector of independent uniform random variables on $[0, 1]$. We simulate each process many times for various T between 10 and 10000 (immediately stopping each trial upon reaching T), retaining the same v and A except when changing N.

We present the estimation error, as well as several predictors, in Figure 3.2. Specifically, we compare the squared-Frobenius error to the corresponding trace-of-inverse of (3.26) multiplied by various intensity measures (empirically estimated by averaging across identical trials and summed over N). We also compare against the Cramér-Rao bound $\sum_{i \in \mathbb{Z}_N} \operatorname{Tr}((\mathfrak{I}^{(i)})^{-1})$, which we estimate, empirically, from the observations using (3.4).

In examining Figure 3.2, we see that the performance typically converges to the Cramér-

Rao bound relatively quickly, as T grows. This value is comfortably nested between the minimum and maximum intensities times the expected Gram matrix, as dictated by (3.9). Further, in these instances, the asymptotic error (dictated by the Fisher information) is related to the expected Gram matrix by a factor somewhat-near to the the average intensity, and this correspondence becomes somewhat more accurate with increasing N. While we present only a single set of model parameters for each N, additional simulations showed these results to be representative of the general trends.

We remark that the empirical intensities λ_{mean} and λ_{max} are smaller-than-expected for small T because we initialize these processes with $y_0 = 0$, rather than allowing the process to "warm-up" before starting our observations. The difference between λ_{mean} and $\|Bv\|_1$ gives an idea as to the size of this discrepancy. The minimum intensity λ_{min} is unaffected because it is determined by v. While the minimum and average intensities can be reliably estimated, the maximum must be bounded using techniques such as those in [33].

3.7.2 Estimation of excitation

Next, we present a simulation based on estimation of parameters of first-order LARPPs. For our systems, we set $N = 5$ and $T \in [50, 10000]$ to be log-uniform (e.g., the chance of a value in the interval $[100, 200]$ is equal to the chance of a value in the interval $[200, 400]$). We then choose the base intensity vector v to have independent uniformly-distributed entries in the interval $[0, 1]$. We choose the excitation matrix A by initially generating a matrix of independent exponential random variables, then scaling the entries so that the spectral radius is a uniform random variable on the range $[0, 0.95]$. This provides us with a diversity of different systems. We simulate one realization of each system, assume that the interval T and base intensity vector v are known, and use maximum likelihood estimation (3.3) to provide an estimate of the true excitation $\widehat{A}$.

We will use our conjecture from Section 3.6.2, which was supported by simulations in Section 3.7.1, that the mean intensity is useful for predicting the estimation error. For each

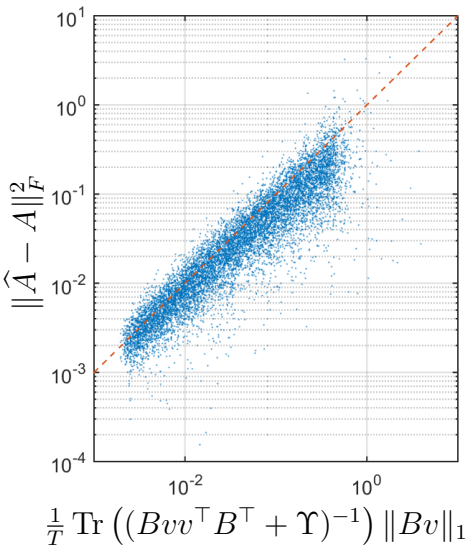

Figure 3.3: Scatterplot of error predicted by an approximation of the presented analysis versus the error of the maximum likelihood estimate $\widehat{A}$

system, we compute the quantity

$$\frac{1}{T}\,\mathrm{Tr}\left((Bvv^\top B^\top + \Upsilon)^{-1}\right)\|Bv\|_1,$$

the trace of the inverse of the expected Gram matrix times the expected average intensity of the process. By using the bounds (3.9) and replacing the extreme values of the intensities with the averages (resulting in the $\|Bv\|_1$ term), our theory suggests that this should approximately reflect to the mean-squared Frobenius error $\|\widehat{A} - A\|_F^2$ of the maximum likelihood estimate. We then compare this number to the actual squared Frobenius error of the maximum likelihood estimate. Because the maximum likelihood estimator is asymptotically efficient, we expect these values to be in approximate agreement.

A scatterplot of our predicted and actual error for single realizations of 10000 systems, each designed according to the above scheme, is presented in Figure 3.3. On average, the ratio between our predicted and the actual error was 1.97 or 0.77 (depending on which term takes the numerator of the quotient). In other words, the error we predicted by our coarse approximations exceeded the true error by about a factor of two, or the true error

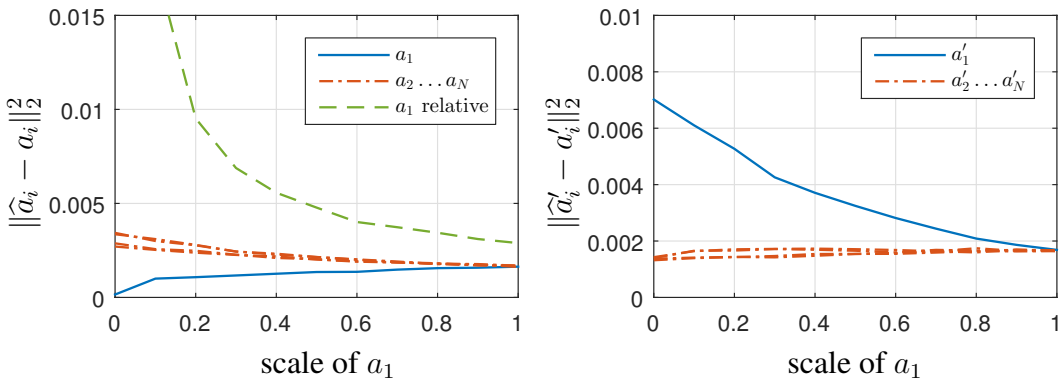

Figure 3.4: How the estimation of the rows (left) and columns (right) of the excitation matrix are affected by scaling one row of the excitation matrix

was smaller than our predicted error by about a quarter.

3.7.3 Interplay between subprocesses

Although (3.7) dictates that we can control the error of A one row at a time, rather than the entire matrix (as we have been doing), the Fisher and Gram matrices introduce interplay between the rows. In this section, we fix a LARPP with $N = 5$ to have $v = [1, 1, 1, 1, 1]^\top$ and a circulant-structured A with first row $a_1 = [0, 0.4, 0, 0, 0.4]$, leading to $\rho(A) = 0.8$. We then scale the first row linearly between 0 and this value (with 1 representing the specified value) while holding v and all other rows constant. We simulate many instances of this process at different scales and $T = 1000$ and present the error in estimating each row or column of A as the scale factor changes. We measure this error as $\|\widehat{a}_i - a_i\|_2^2$ for the rows of A and $\|\widehat{a}'_i - a'_i\|_2^2$ for its columns (where a_i represents row i of A and a'_i represents column i of A).

We present the results of this procedure in Figure 3.4. We see that, as a row of A grows, the absolute error of estimating that row will typically grow while the relative error will decrease. As the row grows, the subprocess will become more active and experience more events. This elevated activity will provide more events from-which to estimate excitation, improving relative performance. Absolute performance is degraded as the correct estimate moves away from zero. The other rows (with nonscaling coefficients) will see a small

improvement in performance as a result of increased activity induced by a more-active first subprocess. The non-adjacent rows (3 and 4) perform slightly better than the adjacent rows (2 and 5) for small scales, as the adjacent rows have less activity.

A column of A represent how much influence a process has on all others. Increasing the first row of A will improve estimates of the first column a'_1 (e.g., how much the first subprocess affects others) due to increased activity. However, the elevated activity will also result in more ambiguity in estimating other columns. Again, at small scales of a_1, the non-adjacent columns perform marginally better than adjacent ones because of the differing activity levels.

From this simulation, we see that there is a counterplay between the activity of different subprocesses and how well different parts of A are estimated. As activity increases, more events are available to improve the estimation of parameters. However, increasing activity can also increase the ambiguity that we experience in attributing activity to a connection in a network. In simulation, we saw that increasing the excitation parameters of one row of A resulted in improved performance for that column, improved relative (though not absolute) performance for that row, sightly improved performance in estimating other rows, and slightly degraded performance in estimating other columns. This aligns with the conjecture from (3.28) that performance is best when all subprocesses have similar activity levels.

3.8 Estimating unknown kernels

In Section 3.2, we laid out the simple scenario where $\mu_i(t)$ and $\phi(t)$ are known. However, we may sometimes wish to use different kernels for different relationships, specifying distinct a $\phi_{ij}(t)$ for each $\gamma_{ij}(t)$. Further, in some cases $\mu_i(t)$ and $\phi_{ij}(t)$ are unknown and must be estimated in parallel with A [46]. The result of Theorem 3 can be easily modified (requiring a careful accounting of subscripts but the same exact approach) to account for the

more-flexible model

$$\lambda_i(t) = \sum_p w_{ip}\psi_{ip}(t) + \sum_q \sum_{j\in\mathbb{Z}_N} \alpha_{ijq} \sum_{k\in S_j} \phi_{ijq}(t-\tau_k), \tag{3.29}$$

where $\mu_i(t)$ and $\gamma_{ij}(t)$ have been broken up into compositions of multiple kernel functions. By estimating all the parameters $\{w_{ip}\}$ and $\{\alpha_{ijq}\}$, one can estimate $\mu_i(t)$ and $\gamma_{ij}(t)$ in greater detail. If we continue to assume that each $\phi_{ijq}(t)$ is normalized to have unit-integral then the excitation matrix is $A_{ij} = \sum_q \alpha_{ijq}$.

The Gram matrix is still formed by the expected cross-correlations of our kernel functions, but now we have (adjusting our notation from Section 3.5 to index Φ by functions instead of indices)

$$\Phi_{\psi_{ip}\psi_{iq}} = \int \psi_{ip}(t)\psi_{iq}(t)dt \tag{3.30}$$

$$\Phi_{\psi_{ip}\phi_{imq}} = \Phi_{\phi_{imq}\psi_{ip}} = \mathbb{E}\left(\sum_{k\in\mathbb{S}_m} \int \psi_{ip}(t)\phi_{imq}(t-\tau_k)dt\right) \tag{3.31}$$

$$\Phi_{\phi_{imp}\phi_{inq}} = \mathbb{E}\left(\sum_{k\in\mathbb{S}_m}\sum_{\ell\in\mathbb{S}_n} \int \phi_{imp}(t-\tau_k)\phi_{inq}(t-\tau_\ell)dt\right), \tag{3.32}$$

whereas we previously only had expressions of the form (3.32). The value (3.32) can be computed using (3.16), exactly as before except that we replace $R_\phi(\tau)$ with the corresponding cross-correlation $R_{\phi_{ijp}\phi_{ijq}}(\tau) = \int \phi_{imp}(t)\phi_{inq}(t-\tau)dt$. Fortunately, (3.30) and (3.31) are simpler. The value of (3.30) is nonrandom and so it can be computed by simple evaluation. The expectation of (3.31) depends on only one random variable at a time, and can be computed to be

$$\Phi_{\psi_{ip}\phi_{imq}} = \int \left(\int \psi_{ip}(t)\phi_{imq}(t-\tau)dt\right)\eta_m(\tau)d\tau.$$

Bounds of the form (3.9) can be computed using this new Gram matrix. However, the larger number of free parameters will make recovery more challenging. Regularization can

mitigate the increased dimensionality, in practice, although the theory presented here does not account for these benefits.

3.9 Discussion

We have provided a calculation of the expected Gram matrix that is induced by attempting to estimate the excitation matrix of a Hawkes process. By its relation to the Fisher information matrix, this allowed us to compute the Cramér-Rao lower bound for the estimation error. The Cramér-Rao bound was used to characterize the asymptotic performance of the maximum likelihood estimator.

Most of our detailed discussion focused on the case of first-order linear autoregressive Poisson processes with stationary statistics, in order to obtain a tractable analysis. We anticipate that the general behaviors observed in this special case may be reflected among Hawkes processes that utilize excitation kernels of modest extent (which covers most applications).

Our original intent was to explore what structures in the excitation matrix made the recovery task easier or more difficult. However, in Section 3.6.3 we discussed how the general structure of the excitation matrix plays a relatively small role compared to other system properties.

In Chapter 2, we observed that the *empirical* Gram matrix could play an important role in providing nonasymptotic error bounds for Poisson processes. By analogy, it is possible that [23], which provides concentration inequalities for the entries of the empirical Gram matrix, could be helpful to establishing nonasymptotic estimation error bounds for Hawkes processes. Better results would likely be achieved by concentration inequalities that directly address the eigenvalues of the empirical Gram matrix, which remain unexplored.

It is widely observed that incorporating sparse priors into estimation procedures can improve performance, and point processes are no exception [32]. Especially in large networks, sparsity is an extremely common characteristic of real-world systems. While our

analysis failed to present a case for sparsity-based gains, empirical findings suggest that they can be substantial [11], in holding with broader trends on sparse parameter estimation. Further work into this problem will be necessary to yield theory that considers the benefits of sparse priors.

CHAPTER 4

HAWKES MODELS FOR TELECOMMUNICATIONS

A fundamental problem in network monitoring consists of characterizing and analyzing the structure of a wireless network. There are a variety of techniques for approaching these problems as an observer *within* the network, but in many important applications we are limited to the role of a passive observer *outside* the network, meaning that we cannot directly observe the content of the network traffic (message content, routing information, etc.). This may arise in cases where the network traffic is encrypted or otherwise unavailable, or in cases where there is simply too much traffic to process. These scenarios are particularly common in signals intelligence and electronic warfare applications, but may arise in many other application areas as well.

Our focus in this chapter is on how we can solve problems, such as learning the network topology and detecting changes to the existing topology, under the assumption that the only information we can observe is when each particular transmitter in the network initiates a transmission. While learning from this limited source of data is challenging, we will see that a significant amount of information can be extracted from the data by exploiting the fact that communication is typically *reciprocal*. That is, a transmission from a particular transmitter is likely to cause other transmissions in response (such as a return message, acknowledgment, packet forwarding, etc.). With this assumption, we can use the co-occurrence of transmissions from different transmitters to infer their relationships by modeling our data with a multivariate Hawkes process. Note that this approach relies on the assumption that we are able to accurately determine who is transmitting at any given time. In practice, this can be achieved through geolocation, specific emitter identification, and other content-agnostic features.

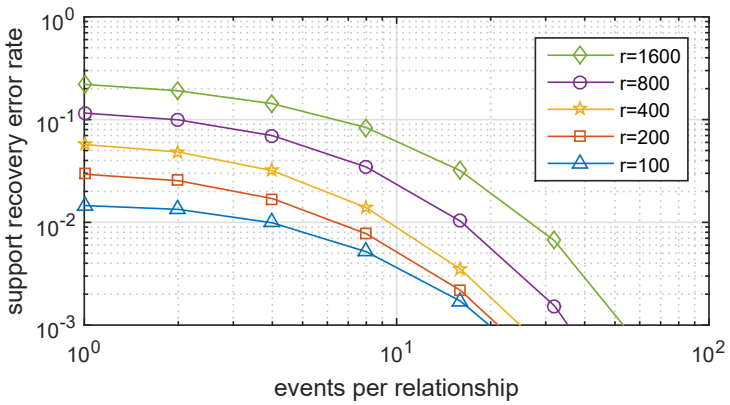

Figure 4.1: Support recovery rate as a function of the number of events observed per relationship. Each line denotes a different number of relationships (nonzeros) in the 80×80 influence matrix.

4.1 Parameter inference accuracy

An important question when performing inference concerns to the number of events we must observe in order to learn the parameters to a given accuracy. The motivation of Chapter 2 and Chapter 3 is to address similar problems that may provide useful insights from mathematical analyses. While they fall short of conclusively addressing this question, they do hint that difficulty should be roughly linear in the number of connections (rather than free parameters) and that a diversity of behaviors provides for slightly easier recovery. Here we will address the question from intuitive and empirical perspectives.

Clearly, in order to discover a connection between two elements in the network (i.e., a nonzero entry in A), we must see it used at least once. Thus, if there are r nonzeros in A, then the *coupon collector* problem suggests we will require at least $O(r \log r)$ events to observe all connections "in action" at least once. More realistically, we will want to see connections used several times so that we can deal with ambiguous cases and avoid false-positives. In other words, we expect the number of events needed to accurately recover the locations of the nonzeros in A to be slightly super-linear in the number of nonzeros.

Figure 4.1, borrowed from [47], provides evidence of this behavior. It considers an 80 node network where bidirectional interconnections are randomly added such that the

expected total number of connections in the network is r. Each connection has the same strength and that magnitude is chosen so that the Hawkes model is subcritical ($\rho(A) < 1$). Events are generated using a multivariate Hawkes process with these parameters and we then compute the maximum likelihood estimates $\widehat{\mu}$ and $\widehat{A}$. The resulting entries of $\widehat{A}$ are thresholded before computing the error rate in estimating the locations of the nonzeros (the *support recovery error rate*). This is repeated for many trials for each r, and number of observations and the error rates are averaged and reported in the figure. The horizontal axis is displayed in terms of the number of events observed divided by r. The results of the simulation are consistent with our intuition that the number of events is weakly super-linear in r (something like $O(r \log r)$).

There are a number of additional factors that likely play into the scaling, as well. As the overall event rate increases, greater ambiguity occurs as it becomes more difficult to determine which subprocess is responsible for exciting another. This can eventually lead to a dramatic increase in the number of event observations required to perform inference. Other issues that can affect this requirement include the dynamic range of the entries of A and mismatch between the Hawkes process and the actual underlying process. However, depending on the actual model, this mismatch can sometimes make inference easier rather than more difficult.

4.2 Application to wireless networks

In practice, we do not expect the data in a real-world network to truly follow a Hawkes process. For example, there will typically be additional structure in the data not captured by the Hawkes model. However, a Hawkes process does capture the reciprocal nature of the interactions we would expect to observe and so, despite some degree of model mismatch, it can be useful as an inference tool. Here we provide a brief demonstration of the use of Hawkes processes applied to realistic network communication data.

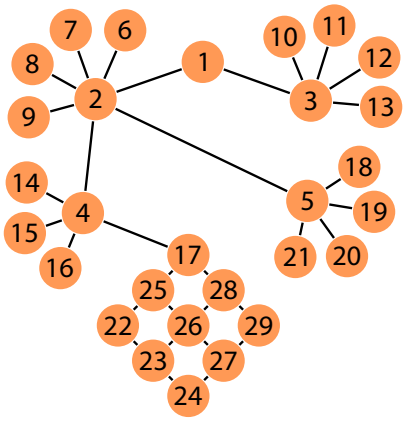

Figure 4.2: Arrangement of simulated EMANE network. Some links are not utilized in the simulation.

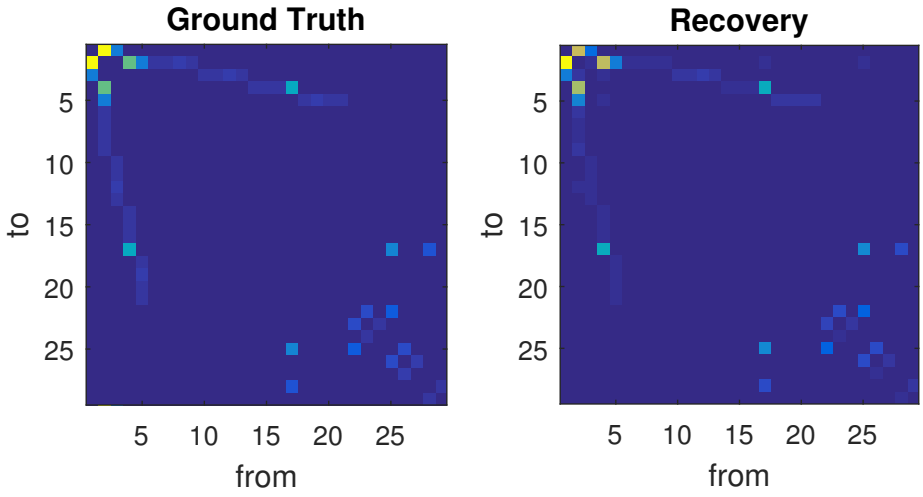

Figure 4.3: Ground truth and recovered influence of EMANE-simulated data.

The data we will use is a trace created by the EMANE network emulator.[1] The network used consists of 29 interconnected nodes, arranged as in Figure 4.2, transmitting a total of over 2.2 million packets. Packets are generated and then propagated along the network from node to node to reach their destination. We then strip the trace of all routing information, except for the transmitter of each packet, to provide a simulation of the kind of data we would be able to observe in a typical wireless network surveillance scenario.

We then compute the maximum likelihood Hawkes parameters. We use an exponential decay function kernel for $\gamma(t)$, though typically the exact shape of the kernel has a limited

[1] http://www.nrl.navy.mil/itd/ncs/products/emane

impact relative to more general features such as its duration. We find it useful to impose the constraint that $A_{ii} = 0$, since we have no reason to believe that a radio will transmit information in response to its own previous transmissions.

The ground truth and the recovered influence are shown in Figure 4.3. While a relatively small number of false-positives have persisted and some of the weaker links have been missed, all of the strong connections within the network have been recovered. We emphasize that even though the data was not generated according to a Hawkes process, this model was able to exploit the reciprocity in the network to discover most of the connections.

4.3 Detecting changes in the network

A benefit of using Hawkes processes to model data is that it provides a means to asses the plausibility of an observation. Using this, we can determine whether sets of data are drawn from the same distribution. In the specific context of wireless networks, we can determine if two sets of observations are plausible under the same network topology. This is useful if we wish to recognize when the topology of the network changes.

Towards this end, an effective measure of deviation between two sets of data is the log-likelihood distance. We define $\{\widehat{\mu}_{1i}, \widehat{a}_{1i}\}$ and $\{\widehat{\mu}_{2i}, \widehat{a}_{2i}\}$ to be the maximum likelihood parameters for subprocess i for the sets of observations $\{\tau\}_1$ and $\{\tau\}_2$, respectively. We then define the deviation of observations $\{\tau\}_2$ from the model suggested by $\{\tau\}_1$ (with regard to subprocess i) to be

$$d_i(\{\tau\}_1, \{\tau\}_2) = \frac{\mathcal{L}(\{\tau\}_2 | \widehat{\mu}_{1i}, \widehat{a}_{1i}) - \mathcal{L}(\{\tau\}_2 | \widehat{\mu}_{2i}, \widehat{a}_{2i})}{|\{\tau\}_2|}. \tag{4.1}$$

This expression is always nonnegative because $\{\widehat{\mu}_{2i}, \widehat{a}_{2i}\}$ are the likelihood maximizers $\{\tau\}_2$. The denominator serves to appropriately scale the deviation to account for the property that the negative log-likelihood scales linearly with the number of events. This

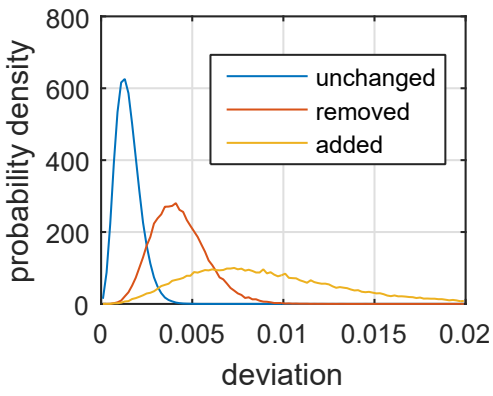

Figure 4.4: Probability density functions showing how likelihood deviation changes when single connections are added to or removed from nodes in a network.

quantity is asymmetric, but can be symmetrized if necessary by defining $d_i'(\{\tau\}_1, \{\tau\}_2) = d_i(\{\tau\}_1, \{\tau\}_2) + d_i(\{\tau\}_2, \{\tau\}_1)$.

One possible use of this metric is to aid in detecting changes in network topology. As an example, we consider a 50-node network modeled by a multivariate Hawkes process where the matrix A is binary-valued (scaled to maintain stability) with 200 nonzeros that are uniformly selected at random. We assume that we are given a long period of observations to estimate μ and A and we then modify 10% of the nodes by adding an additional connection (i.e., adding an additional nonzero to the corresponding row of A) and another (different) 10% by subtracting a connection. We simulate 5000 events (25 per connection) using these new parameters and we call the resulting observations $\{\tau\}_2$. With a slight abuse of the notation in (4.1), we calculate $d_i(\{\mu_i, a_i\}, \{\tau_2\})$, the change likelihood of the observed events when using the maximum likelihood parameters versus the parameters of our base model. We categorize nodes as either unchanged (having the same links in both the base and modified processes) or as nodes with a removed or added connection, and examine the distribution of the d_i's for each group.

Empirical distributions of the d_i's, estimated from thousands of trials, are presented in Figure 4.4. Observe that it is relatively easy to recognize when a new connection is added (the "added" distribution has little overlap with the "unchanged" distribution) but

it is somewhat more challenging to recognize when a connection is removed. Intuitively, this phenomenon can be explained by the fact that new connections can be identified as soon as unexpected transmissions occur, whereas a missing connection requires the more subtle detection of the absence of transmissions that we expected to see. In both cases, the modified distributions deviate further from the unchanged distribution as more events are observed, making the discrimination easier.

4.4 Incorporating additional structure via marks

The Hawkes process model we have considered up to this point captures the "conversational" aspect of typical network interactions. However, it still may be a poor representation of actual behavior in some specific networks. For example, cell phones do not exhibit a back-and-forth transmission pattern (like a push-to-talk or packet radio network) but instead establish one enduring transmission (or continuous string of transmissions) for the entire interaction. We may further expect that transmission lengths will be comparable between two interacting nodes. The simple multivariate Hawkes process described above cannot model this additional structure.

Fortunately, we can incorporate such details into the model via the addition of *marks* – additional information corresponding to each event – to a Hawkes process. Marks must be drawn from some distribution and can depend on other event times and marks. As discussed in Section 1.3.5, the multivariate Hawkes process can be viewed as a marked Hawkes process in one dimension.

As a more elaborate example, we will describe a model that better resembles cellular traffic. Let us say that for every phone we can observe the start and end times of all calls made, but not with whom they are speaking.[2] For this example, let us assume that calls have a duration distributed with probability density function $\psi(t)$, that a phone that receives a call will answer within 20 seconds, and that both phones will terminate the call within 5

[2]We could track base stations as well, but they convey the same information (in terms of what we observe) so we will ignore them in this discussion.

seconds of each other. We will use τ_k to indicate the start time of a call and θ_k to denote the end time. A CIF that incorporates this assumption might be

$$\lambda_i(t, u|\tau_S, \theta_S) = \mu_i\psi(u-t) + \sum_j \frac{A_{ij}}{20} \sum_{k \in S_j} \mathcal{I}_{(0<t-\tau_k \leq 20)} \mathcal{I}_{(-5 \leq u-\theta_k \leq 5)} \tag{4.2}$$

where $\mathcal{I}_{(\cdot)}$ is 1 when the argument is true and 0 otherwise.

This model, while much more realistic than a standard Hawkes process, still has some limitations. It theoretically allows for many phones to engage in the same call and allows individual phones to be engaged in multiple calls at once (or the same call multiple times, even). It doesn't account for missed calls or a number of other scenarios, either. But while this model may be lacking in a generative sense, when we use it as an inferential model, we do not necessarily need to be overly concerned by such cases since they do not typically arise in the actual data. What this model *does* impose (in the context of inference) is the assumption that if two phones do not have start and end times closely matched then they were very unlikely to be involved in the same call. In simulations we have observed that by incorporating this additional assumption we can dramatically improve the quality of our inferences.

This illustrates that while the base model for a Hawkes process may not perfectly match some applications, marks provide a powerful technique to extend it. With creativity, one can produce models that impose a variety of additional structural assumptions.

4.5 Discussion

We have demonstrated a number of techniques and observations that may be useful for modeling behavior patterns in wireless networks with Hawkes processes. This convenient statistical model captures the behaviors common in communication networks and allows for useful inferences to be made from limited information.

There are a many additional variations not discussed here. For example, there is no need

to restrict ourselves to a single kernel or a single influence matrix, additional terms (involving different kernels and/or matrices) can be added without affecting the convexity of the inference program. This possibility was discussed in Section 3.8. These additional kernels and coefficients can dramatically enhance the representative power of Hawkes models. It is even possible to learn kernels from the data by inferring different weights at different delays [46]. In Chapter 5, we discuss a relevant extension where we use the Hawkes model to not only detect links within the network, but to learn the flow of information between nodes. This is especially relevant to telecommunication networks, which may utilize relay nodes to achieve full connectivity.

CHAPTER 5

EXCITATION ATTRIBUTION IN HAWKES PROCESSES

A fundamental problem in the analysis of complex networks is determining how information or other quantities flow through the network. For example, in a communication network we might be interested in who is communicating with whom; in a network of neurons we may wish to know how certain neurons affect other neurons in the network; or in a financial network we might want to identify the end parties of transactions that pass through (possibly many) intermediaries.

When we can observe interactions between nodes in the network in detail, there are a number of techniques that can help to identify this kind of structure. However, in many cases we may have only limited information about what is happening. For example, in a wireless network we might be able to identify the source of a transmission but not the destination. In many other communication/social networks, privacy concerns and other practical limitations might preclude knowledge of the intended recipient of an interaction. However, the timings of events (e.g., transmissions in a wireless network) in the network are often easier to obtain. Furthermore, this information can be sufficient to learn about the network if we assume that events at one node are likely to induce a response at some other node.

This can be formalized when we model our system as a multivariate Hawkes process. While fitting a Hawkes process to a collection of recorded events can provide insight to the structure of the network, it is largely limited to discovering the direct connections. If nodes act as intermediaries between others, the meaningful interactions in a network can become obfuscated. For example, packets in wireless ad hoc networks may be forwarded multiple times to reach their final destination. It may be of greater interest to know the source and destination of the packet rather than the topology of the network. A Hawkes model can be

used to determine which nodes are connected, but does not directly reveal this source/sink information.

Here, we explore novel techniques based that aim to uncover the relationship between events more explicitly in the Hawkes model. Recognizing relationships between events (rather than nodes) can enable us to discover more complex interactions within a network, something not provided by existing approaches based on Hawkes processes.

5.1 Inter-event influence

The Hawkes model model provides a way to estimate how much influence *nodes* have on each other in a network, but another important question concerns how much influence individual *events* exert. Specifically, we might wish to know which events are responsible for causing which other events in a sample. To address this question, we will assume that events have a single cause – events can be generated either spontaneously (for reasons internal to the node) or can be a result of a *single* previously-observed event. We used this as an interpretation in Chapter 3 within the context of a generative model, which was permissible thanks to Poisson superposition. However, here we must accept it as explicitly true in a real sense for our following interpretations to hold.

The intensity function of a Hawkes process dictates the rate of event occurrence through a compositional structure that adds the contributions of multiple terms, as in (1.12). Let θ_k be the node that created the event at time τ_k (the i such that $k \in S_i$) and consider $\gamma_{i\theta_k}(t-\tau_k)$, which is just one component of the overall intensity function. $\gamma_{i\theta_k}(t-\tau_k)$ tells us how much intensity event k contributes to process i at any time t. Because intensity gives the probability of occurrence within a differential time interval, we can make probabilistic statements about the cause of an event that occurs at a specific time. By evaluating $\gamma_{\theta_\ell\theta_k}(\tau_\ell - \tau_k)$ we can quantify the contribution of event k toward generating an event at time τ_ℓ (on process

θ_ℓ). The probability that event k is responsible for event ℓ is

$$\mathbb{P}(k \to \ell) = \frac{\gamma_{\theta_\ell \theta_k}(\tau_\ell - \tau_k)}{\lambda_{\theta_\ell}(\tau_\ell)}. \tag{5.1}$$

We can organize these probabilities into an *attribution matrix* E where $E_{k\ell} = \mathbb{P}(k \to \ell)$. Sorting the events by arrival order ensures that this matrix is upper triangular; because $\gamma_{ij}(t)$ is strictly causal, there is no possibility that events are caused by later or concurrent events (including themselves). This means our event attribution probabilities form a (weighted) directed, acyclic graph. If the $\gamma_{ij}(t)$ are time-localized (zero beyond some modest value of t) then the attribution matrix will also be banded.

We can also use this scheme to consider chains of potentially connected events. For example, the probability that event a caused event b which in turn resulted in event c is $\mathbb{P}(a \to b) \cdot \mathbb{P}(b \to c)$, and the same can be done for arbitrarily long candidate sequences. We can compute the probability that an event is an innovation (i.e. not the result of a previous event, but rather the start of a new chain) as

$$\mathbb{P}(\varnothing \to \ell) = \frac{\mu_{\theta_\ell}(\tau_\ell)}{\lambda_{\theta_\ell}(\tau_\ell)}, \tag{5.2}$$

or the probability that an event k is terminal (does not result in another event) as

$$\mathbb{P}(k \to \varnothing) = \prod_\ell (1 - \mathbb{P}(k \to \ell)). \tag{5.3}$$

One can use these quantities to identify the endpoints of a chain of related events.

5.2 Event chains

If we wish to find the most likely chain of events that includes a particular event then a simple dynamic program, such as the Dijkstra algorithm [48], can be used to do so. In some contexts, this may be all we care about. However, given the significant ambiguity

that often arises in our model, the "most probable" chain can still be very unlikely. We will see later that it can be beneficial to consider additional candidates to gain an effective understanding of the network dynamics.

Even if E is banded and acyclic, the number of possible chains in the graph is still exponential in the total number of events. But if we only consider chains with at least some minimum probability, we usually observe a dramatic reduction (to roughly linear in the number of events). This can still be a large number, however, so it will be worth examining an efficient way to explore them all.

We want to find all sequences of events (paths) $W = \{w_1 \dots w_L\}$ such that $\mathbb{P}(W) \geq \alpha$, where

$$\mathbb{P}(W) = \prod_{i=1}^{L-1} E_{w_i, w_{i+1}}. \tag{5.4}$$

To do this, we will use a graph traversal algorithm we will call *Low Cost Paths for Acyclic Graphs* (LCPAG). This dynamic program addresses the *all simple paths* problem [49] with modifications such that it prunes paths with excessive cost and does not work on graphs with cycles. To maintain consistency with other graph search algorithms, which traditionally use additive rather than multiplicative cost, we will instead present an algorithm to find all paths W such that $C_W \leq \alpha'$ where

$$C_W = \sum_{i=1}^{L-1} E'_{w_i, w_{i+1}}. \tag{5.5}$$

This is equivalent to our problem if we use $E'_{ij} = -\log E_{ij}$ and $\alpha' = -\log \alpha$.

The algorithm is presented in Algorithm 2. The idea is that if we know all the paths (and their costs) from a k_i to k_{stop} then finding a path to k_i is just as good as finding k_{stop}. We begin our search from k_{start}, where we ask each of its neighbors for every path (and associated cost) they know that goes through them and reaches k_{stop}. If a neighbor's path's cost plus its cost from k_{start} is less than α', we extend that path by connecting k_{start} (and record the new cost). Now we know every way to reach k_{stop} from k_{start}.

Algorithm 2 Low Cost Paths for Acyclic Graphs

inputs: $E' \in [0,\infty)^{n\times n}$, $k_{\text{start}} \in \{1\dots n\}$, $k_{\text{stop}} \in \{1\dots n\}$, $\alpha' \in [0,\infty)$
initialize: $R^{\text{cost}}_{k_{\text{stop}},1} = 0$, $R^{\text{node}}_{k_{\text{stop}},1} = k_{\text{stop}}$, $R^{\text{next}}_{k_{\text{stop}},1} = \varnothing$,
$u_{\{1\dots n\}\setminus k_{\text{stop}}} = \text{true}$, $u_{k_{\text{stop}}} = \text{false}$, $n^{\text{path}}_{k_{\text{stop}}} = 1$
call: TRAVERSE(k_{start})
for $i \in \{1\dots n^{\text{path}}_{k_{\text{start}}}\}$ **do**
  $r = R_{k_{\text{start}}i}$, $C_i = r^{\text{cost}}$, $\ell = 0$
  while $r \neq \varnothing$ **do**
    $\ell = \ell + 1$, $P_{i\ell} = r^{\text{node}}$, $r = r^{\text{next}}$
return P, C

SUBROUTINE: TRAVERSE(k)
initialize: $\ell = 0$
for i s.t. $E'_{ki} \leq \alpha'$ **do**
  if u_i **then**
    recurse: TRAVERSE(i)
  for $j \in \{1\dots n^{\text{path}}_i\}$ **do**
    if $E'_{ki} + R^{\text{cost}}_{ij} \leq \alpha'$ **then**
      $\ell = \ell + 1$, $R^{\text{cost}}_{k\ell} = E'_{ki} + R^{\text{cost}}_{ij}$, $R^{\text{node}}_{k\ell} = k$, $R^{\text{next}}_{k\ell} = R_{ij}$
$n^{\text{path}}_k = \ell$, $u_k = \text{false}$
return

Of course, k_{start}'s neighbors may not know how to reach k_{stop} and, if they don't, must ask their neighbors before they can answer. We recurse the function for every node (or at least those that could potentially be along viable routes) that doesn't know the paths it can take to reach k_{stop}. Once a node knows all the paths to the end, we mark it (set u_k to false in the algorithm) and use its saved result instead of recomputing. The constraint that the graph is acyclic ensures that we can finish evaluating a node completely before we evaluate its parent. We can follow R like a linked list (though it is actually a reversed-tree or funnel), inspecting the cost from the head, to see what events are in a possible chain. The final stage of the algorithm (the part following the subroutine call) simply does this to provide a listing of every path.

We can incorporate the innovation and termination probabilities by adding two dummy vertices. One dummy vertex connects to every event with the corresponding innovation probability and the other is connected from every event using the termination probability. By finding all plausible paths from the innovation dummy to the termination dummy, we

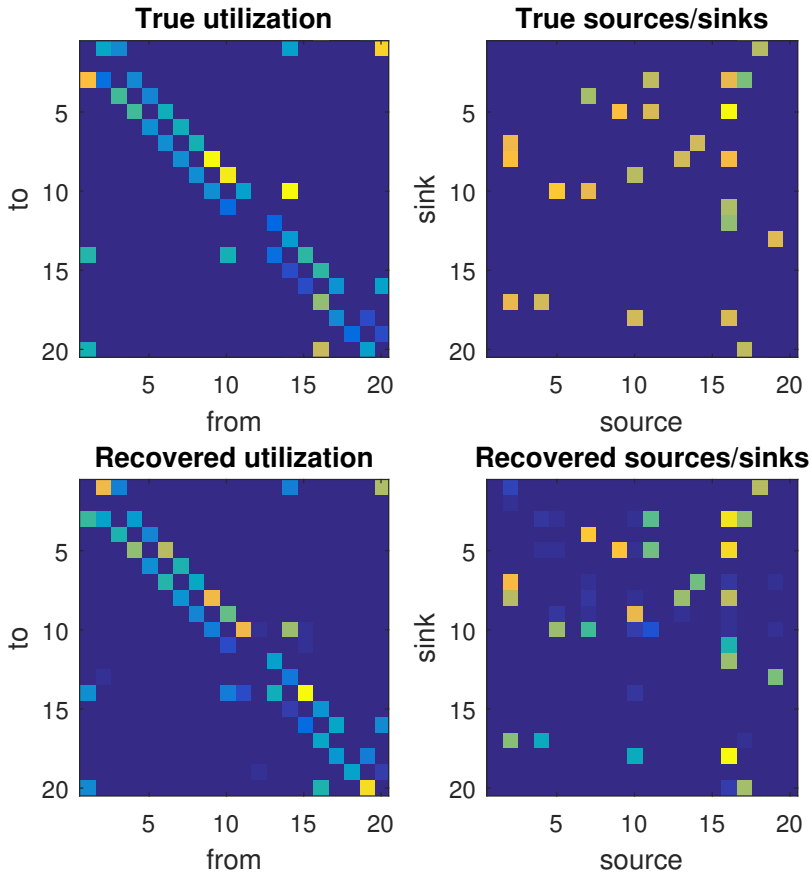

Figure 5.1: Ground truth (top) and recovered (bottom) link utilization (left) and source/sink pairs (right, measured as in (5.6)) for example simulation using $\alpha = 50\%$. Yellow (bright) values indicate stronger connections. Though the Hawkes model accurately infers the utilization parameters, it may not reflect the actual behavior of the network.

have found every possible chain of events with probability α or greater.

5.3 Tracking the flow of information

The LCPAG algorithm allows us to group events in a Hawkes process into probable chains of events. This can allow us to track the flow of information in a network. Identifying common information pathways (for example, identifying pairs of nodes which are communicating with each other) in a complex network is a different problem than identifying the local structure of a network. While multivariate Hawkes models can be effective in learning local network structure, they do not directly provide much insight into the common

information pathways. However, we will see below that the techniques described above can effectively leverage the Hawkes model to learn this deeper structure.

We will consider a simulated network where messages are relayed from one node to another in order to reach their destination. These messages are passed between sets of sources and sinks (destinations). The network begins with 20 nodes that connect in a ring and then adds additional bidirectional connections between pairs of nodes with a 3% probability per pair. We then consider all pairs of nodes and assign them to be a source/sink pair with 5% probability. A source in a pair will send messages to the matching sink by using intermediate nodes (if necessary). When a node receives a message with a destination other than itself, it will pass it to a neighbor that is closer to the sink. Additionally, when a message arrives at the sink, we assume that the sink responds with a single message (to which no one will respond). In many scenarios the sink might reply by sending a message back to the original source, but since single (isolated) transactions between source/sink pairs likely represents the most challenging case, we restrict our attention to this setting. We observe 2000 source/sink transactions (event chains) in our simulation, approximately 105 per source/sink pair.

It is important to note that in order to simulate such a network, we do *not* generate data using a Hawkes process as such a model does not work to deliberately forward a message to a specific destination. While we do not use it for generation, we can choose to fit a Hawkes model to the data. One such realization is shown in Figure 5.1. We observe that although the Hawkes excitation parameters in the matrix A provide an accurate estimate of the local connectivity (and utilization) of the different links in the network, they do not resemble the underlying source/sink structure.

To perform source/sink recognition, we use the LCPAG algorithm to identify likely chains of events. Let $\mathcal{P}_{ij}^{\alpha}$ be the set of all event chains that begin with an event corresponding to node j and end with an event corresponding to node i and have probability at least α. We can estimate the number of event chains beginning with node j and ending with node i

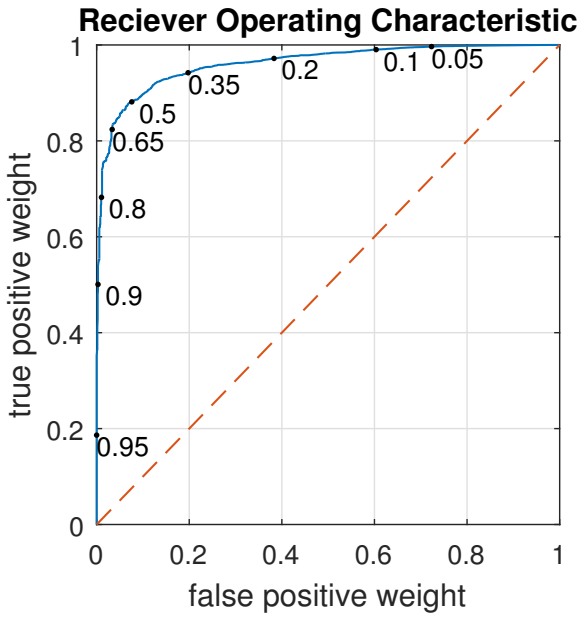

Figure 5.2: Example comparison of hit and false-alarm rate of sum-probabilities ("What fraction of the total weight of correct and incorrect chains do we find?") included when limiting consideration to paths of some minimum probability. The probability thresholds α that result in certain trade-offs are labeled (e.g., $0.5 = 50\%$).

as

$$B_{ij}(\alpha) = \sum_{\eta \in \mathcal{P}_{ij}^{\alpha}} \mathbb{P}(\eta). \tag{5.6}$$

We will find many possible paths, many more than truly exist on the network. One simple way to deemphasize the influence of false paths is to increase α, since we expect the correct paths to typically hold with higher likelihood. This results in a *receiver operating characteristic* (ROC) trade-off. Figure 5.2 shows how the choice of probability threshold determines what fraction of true and false paths we find, weighted by their probability.

This source/sink recognition procedure can be very effective, as seen in Figure 5.1, although it is naturally subject to some degree of inaccuracy. The primary driver of this inaccuracy is ambiguity caused by heavy event traffic; increasing event density makes it more challenging to understand the intent and effect of any individual event. This density can be increased by either increasing traffic volume or having nodes that act as bottlenecks in the network.

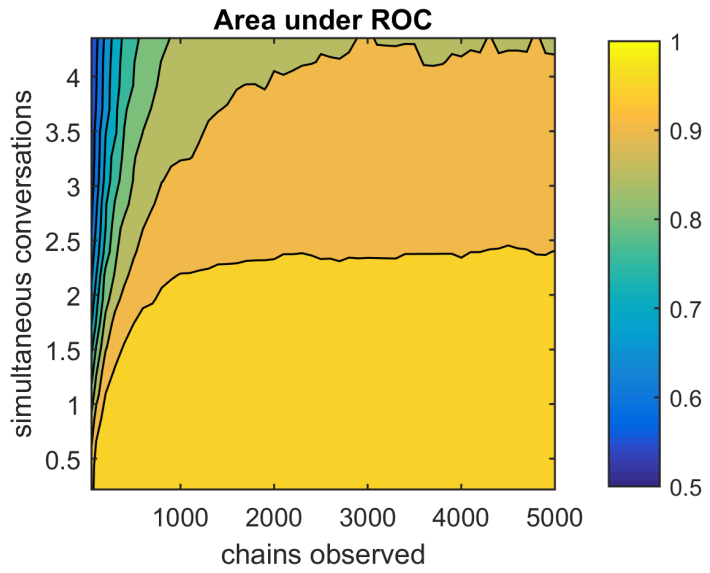

Figure 5.3: Average area under ROC curve when thresholding recovered source/sink pairs. Contours are drawn every 0.05 area.

To visualize the impact of increasing traffic, we simulate many instances of the graph described above. We observe the area under the ROC curve that represents how many correct and incorrect pairs we discover (weighted by their probability) when excluding any source/sink pairs below a given threshold. The averaged results over many trials are shown in Figure 5.3. When the number of observed chains is small, we may be missing many source/sink pairs because they are barely expressed. However, for moderate traffic volumes, we quickly reach a point where we can efficiently identify source/sink pairs. We also note that for any fixed number of observed chains, we see that increasing the traffic density (measured here by the average number of simultaneous conversations) increases ambiguity and limits the performance we can expect to achieve with this method.

5.4 Discussion

In summary, we believe that the simulations described above are a convincing demonstration of the potential for Hawkes models combined with the LCPAG algorithm to identify and track the flow of information in settings where the main source of data is event timings. A limitation of this approach is the drop in performance in high-traffic networks, but we

believe that this could be improved in future work by developing ways to "sparsify" the output of the LCPAG algorithm to eliminate chains of events through unusual sequences of nodes and re-attribute the pruned probability elsewhere.

In [50], a method is presented to modulate a network to produce a desired outcome. The method we describe here could be useful in attempting to evaluate the effectiveness of such an effort, perhaps informing a dynamic control effort.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Theoretical results

The goal of this work was to explore Poisson and Hawkes processes and their use in modeling a variety of systems. We placed specific emphasis on the theoretical questions related to the accuracy with which basis weight parameters of linear Poisson models and excitation parameters of Hawkes processes can be estimated. There were, respectively, addressed in Chapters 2 and Chapter 3.

We addressed Poisson process parameter estimation in Chapter 2. We offered an improvement on existing results that generalized the Poisson counting model to the more-flexible Poisson arrival model. This addressed a major shortcoming of earlier analyses, which suggested *worse* performance as the resolution of observations is improved [22, 32]. This outcome appears to have been inherent to basing the analysis on the Poisson counting process and the consequences of spreading a fixed number of observations over a growing number of bins. Our result, derived by directly considering the (bin-free) Poisson arrival model, avoids this issue.

A similarly-rooted issue manifests in the presence of Poisson processes where the density achieves a high dynamic range. We showed that this can be mitigated by a novel regularization scheme. The scheme we proposed was, however, highly unsatisfactory, in that we artificially reduced the problematic dynamic range by injecting noise into the system. Although the bound we presented improves in the presence of this noise, adding noise to a system rarely improves performance. We conjectured that this may be unnecessary, or may be possible with regularization rather than noising, and supported this conjecture with experiments. However, a formal result to this end remains the subject of future work. An

analysis that organically avoids this shortcoming would be likely to significantly outperform our bound's clumsy handling of the issue.

We were able to provide analogous conclusions to existing work [32] when provided with a sparse prior. However, we were unable to offer computationally-feasible programs to achieve such results. Although there is no indication that our results do not (very-nearly) hold for feasible programs, adapting our theory to avoid computationally-prohibitive constraints will require further development.

In Chapter 3, we provided novel asymptotic results for the estimation of Hawkes excitation parameters. This has substantial utility for attempts to learn network connectivity from event observations. Our results were based on the Cramér-Rao lower bound and a statistical analysis of the basis in which Hawkes estimation occurs. A notable conclusion was that particular architecture of a network plays only a limited role in the complexity of estimation.

However, our Hawkes analysis does not account for sparsity, which we showed in Chapter 4 to have the capacity to substantially affect the accuracy of inference. While the Poisson results of Chapter 2 do not apply to this model, it is nevertheless possible that an analogous result holds. Such an outcome would suggest that an improvement could be had if the Hawkes basis satisfies a restricted isometry or restricted eigenvalue condition. Future work on this topic could yield nonasymptotic Hawkes estimation results that account for sparsity, possibly by exploring the same approach used in Chapter 2.

6.2 Practical results

We showed in Chapter 4 that Hawkes processes can be useful in the context of wireless networks, an area where they previously saw limited use. We presented many adaptions to help them to better model this application. In Chapter 5, we presented a method that uses the branching structure of Hawkes processes to infer relay behavior in networks. This method is particularly valuable in the context of telecommunication networks, where the

Hawkes process captures physical links but this addition is necessary to learn the informational flow within the network.

We also showed that Hawkes processes can capture the salient aspects of networks, even when presented with explicit model mismatch. This is supported by the success of Hawkes processes across a range of real-world datasets. However, it is unclear how reliably the conclusions drawn from Chapter 3 hold for data that is merely similar to a Hawkes process. Further work could characterize which of those conclusions are specific to Hawkes processes and which hold across the broader set of autoregressive systems.

In many applications, the strength of connections is of secondary interest to the simpler question of whether any pair of nodes is connected or not. The work of [51] establishes a method by which it is possible to prune the discovered edges to reduce false-alarms. In [52], a method is discussed for recognizing changes in a dynamic network structure. Although our work, especially Chapter 3, may provide a background for establishing other hypothesis tests for link detection, much work remains on this topic.

REFERENCES

[1] J. Anderson et al. “Weighted least-squares reconstruction methods for positron emission tomography”. In: *IEEE Trans. Medical Imaging* 16.2 (Apr. 1997), pp. 159–165.

[2] A. Singh, D. Wilson, and R. Aufrichtig. “Enhancement of X-ray fluoroscopy images”. In: *Proc. SPIE*. Vol. 1898. 1993, pp. 304–310.

[3] R. Cotter. *Time-of-flight mass spectrometry*. ACS Publications, 1993.

[4] D. Snyder, A. Hammoud, and R. White. “Image recovery from data acquired with a charge-coupled-device camera”. In: *J. Optical Soc. America A* 10.5 (1993), pp. 1014–1023.

[5] R. Willett. “Multiscale analysis of photon-limited astronomical images”. In: *Stat. Challenges in Modern Astronomy*. Vol. 371. 2007, pp. 247–264.

[6] E. Hall and R. Willett. “Foreground and background reconstruction in Poisson video”. In: *Proc. IEEE Int. Conf. Image Processing (ICIP)*. 2013, pp. 2484–2488.

[7] T. LaGrow et al. “Approximating cellular densities from high-resolution neuroanatomical imaging data”. In: *Proc. IEEE Eng. in Med. and Bio. Conf. (EMBC)*. Honolulu, HI, USA, July 2018.

[8] Y. Ogata. “Space-time point-process models for earthquake occurrences”. In: *Ann. Inst. Stat. Math.* 50.2 (1998), pp. 379–402.

[9] K. Zhou, H. Zha, and L. Song. “Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes”. In: *Proc. Int. Conf. Art. Intell. Stat. (AISTATS)*. Scottsdale, AZ, USA, Apr. 2013.

[10] M. Farajtabar et al. “Shaping social activity by incentivizing users”. In: *Proc. Adv. in Neural Processing Systems (NIPS)*. 2014.

[11] M. Moore and M. Davenport. “Analysis of wireless networks using Hawkes processes”. In: *Proc. IEEE Work. Signal Proc. Advances in Wireless Comm. (SPAWC)*. Edinburgh, UK, July 2016.

[12] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. “Inference of functional connectivity in Neurosciences via Hawkes processes”. In: *Proc. IEEE Global Conf. Signal and Information Processing (GlobalSIP)*. Austin, TX, USA, Dec. 2013.

[13] E. Bacry, K. Dayri, and J. F. Muzy. “Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data”. In: *Eur. Phys. J. B* 85.5 (May 2012).

[14] A. Gupta et al. “Hawkes processes for invasive species modeling and management”. In: *arXiv:1712.04386* (2017).

[15] G. Mohler et al. “Self-exciting point process modeling of crime”. In: *J. Amer. Statist. Assoc.* 106.493 (2011), pp. 100–108.

[16] A. Stomakhin, M. Short, and A. Bertozzi. “Reconstruction of missing data in social networks based on temporal patterns of interactions”. In: *Inverse Problems* 27.11 (Nov. 2011), p. 115013.

[17] S. Tench, H. Fry, and P. Gill. “Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army”. In: *European J. Appl. Math.* 27.3 (2016), pp. 377–402.

[18] Z. Harmany, R. Marcia, and R. Willett. “This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms – Theory and Practice”. In: *IEEE Trans. Image Processing* 21.3 (2012), pp. 1084–1096.

[19] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. “Composite self-concordant minimization”. In: *J. Machine Learning Research* 16 (2015), pp. 371–416.

[20] D. Daley and D. Vere-Jones. *An introduction to the theory of point processes*. New York: Springer-Verlag, 1988.

[21] S. Linderman and R. Adams. “Discovering Latent Network Structure in Point Process Data”. In: *J. Machine Learning Research*. 2014, pp. 1413–1421.

[22] M. Raginsky et al. “Compressed Sensing Performance Bounds under Poisson Noise”. In: *IEEE Trans. Signal Processing* 58.8 (2010), pp. 3990–4002.

[23] S. Chen et al. “The multivariate Hawkes process in high dimensions: beyond mutual excitation”. In: *arXiv:1707.04928* (July 2017).

[24] B. Mark, G. Raskutti, and R. Willett. “Network estimation via Poisson autoregressive models”. In: *Proc. IEEE Work. Comput. Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. Dec. 2017.

[25] Y. Bao et al. “Hawkes process modeling of adverse drug reactions with longitudinal observational data”. In: *Proc. Machine Learning for Healthcare Conf.* Vol. 68. Boston, MA, USA, Aug. 2017, pp. 177–190.

[26] Y. Wang et al. “Isotonic Hawkes processes”. In: *Proc. Int. Conf. Machine Learning*. 2016, pp. 2226–2234.

[27] Y. Ogata. “On Lewis’ simulation method for point processes”. In: *IEEE Trans. Inform. Theory* 27.1 (1981), pp. 23–31.

[28] M. Moore, A. Massimino, and M. Davenport. “Randomized multi-pulse time-of-flight mass spectrometry”. In: *Proc. IEEE Work. Comput. Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. Cancun, Mexico, Dec. 2015.

[29] X. Jiang et al. “A data-dependent weighted LASSO under Poisson noise”. In: *arXiv:1509.08892* (2015).

[30] Y. Li and G. Raskutti. “Minimax optimal convex methods for Poisson inverse problems under ℓ_q-ball sparsity”. In: *arXiv:1604.08943* (Apr. 2016).

[31] A. Soni and J. Haupt. “Estimation error guarantees for Poisson denoising with sparse and structured dictionary models”. In: *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*. June 2014, pp. 2002–2006.

[32] M. Rohban, V. Saligrama, and D. Vaziri. “Minimax Optimal Sparse Signal Recovery With Poisson Statistics”. In: *IEEE Trans. Signal Processing* 64.13 (2016), pp. 3495–3508.

[33] N. Hansen, P. Reynaud-Bouret, and V. Rivoirard. “Lasso and probabilistic inequalities for multivariate point processes”. In: *Bernoulli* 21.1 (2015), pp. 83–143.

[34] X. Jiang, G. Raskutti, and R. Willett. “Minimax Optimal Rates for Poisson Inverse Problems with Physical Constraints”. In: *IEEE Trans. Inform. Theory* 61.8 (Aug. 2015), pp. 4458–4474.

[35] M. Davenport et al. “Introduction to Compressed Sensing”. In: *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[36] Y. Ogata. “The asymptotic behaviour of maximum likelihood estimators for stationary point processes”. In: *Ann. Inst. Stat. Math.* 30.1 (Dec. 1978), pp. 243–261.

[37] J. Tropp. “An Introduction to Matrix Concentration Inequalities”. In: *Foundations and Trends in Machine Learning* 8.1-2 (2015).

[38] B. Mark, G. Raskutti, and R. Willett. “Network estimation from point process data”. In: *arXiv:1802.04838* (Feb. 2018).

[39] M. Farajtabar et al. “Fake news mitigation via point process based intervention”. In: *arXiv:1703.07823* (Mar. 2017).

[40] M. Moore and M. Davenport. “Estimation of Poisson arrival processes under linear models”. In: *arXiv:1803.00980* (Mar. 2018).

[41] P. Reynaud-Bouret and E. Roy. “Some non asymptotic tail estimates for Hawkes processes”. In: *Bull. Belg. Math. Soc.* 13.5 (Jan. 2007), pp. 883–896.

[42] H. Watson and F. Galton. “On the probability of the extinction of families”. In: *J. Anthropological Inst. Great Britain and Ireland* 4 (1875), pp. 138–144.

[43] T. Harris. *The Theory of Branching Processes*. Berlin: Springer-Verlag, 1963.

[44] A. Hawkes and D. Oakes. “A cluster process representation of a self-exciting process”. In: *J. Appl. Prob.* 11.3 (1974), pp. 493–503.

[45] M. Moore and M. Davenport. “A Hawkes’ eye view of network information flow”. In: *Proc. IEEE Work. Stat. Signal Processing*. Palma de Mallorca, Spain, June 2016.

[46] K. Zhou, H. Zha, and L. Song. “Learning triggering kernels for multi-dimensional Hawkes processes”. In: *Proc. Int. Conf. Machine Learning*. Atlanta, GA, USA, June 2013.

[47] M. Moore and M. Davenport. “Learning network structure via Hawkes processes”. In: *Proc. Work. Signal Proc. with Adaptive Sparse Structured Representations (SPARS)*. Cambridge, UK, July 2015.

[48] E. Dijkstra. “A note on two problems in connexion with graphs”. In: *Numer. Math.* 1.1 (Dec. 1959).

[49] G. Danielson. “On finding the simple paths and circuits in a graph”. In: *IEEE Trans. Circuit Theory* 15.3 (Sept. 1968).

[50] M. Farajtabar et al. “Multistage campaigning in social networks”. In: *Proc. Adv. in Neural Processing Systems (NIPS)*. 2016, pp. 4718–4726.

[51] S. Chen, D. Witten, and A. Shojaie. “Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process”. In: *Elec. J. Stat.* 11.1 (2017), pp. 1207–1234.

[52] S. Li et al. “Detecting changes in dynamic events over networks”. In: *IEEE Trans. Signal and Info. Processing over Networks* (June 2017), pp. 346–359.